Probability & Measure Theory

Second Edition

Robert B. Ash
Catherine A. Doléans-Dade

Copyright © 2000

Probability and Measure Theory

SECOND EDITION

ROBERT B. ASH

with contributions from
Catherine Doléans-Dade

A Harcourt Science and Technology Company

San Diego San Francisco New York Boston
London Toronto Sydney Tokyo

This book is printed on acid-free paper. ∞

Copyright © 1999 by Academic Press

All rights reserved.
No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

ACADEMIC PRESS
A Harcourt Science and Technology Company
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.apnet.com

ACADEMIC PRESS
24-28 Oval Road, London NW1 7DX, UK
http://www.hbuk.co.uk/ap/

Typeset by Academic Press
28 Wilbraham Road, Hartington, Manchester M20 4LA
http://www.academicpress.com

Library of Congress Cataloging-in-Publication Data
International Standard Book Number: 0-12-045332-1

Printed in the United States of America
99 00 01 02 03 IP 9 8 7 6 5 4 3 2 1

Contents

Preface ...

Summary of Notation ix

1 Fundamentals of Measure and Integration Theory 1

1.1 Introduction 1
1.2 Classes of Sets, and Measures 3
1.3 Extension of Measures 12
1.4 Lebesgue–Stieltjes Measures and Distribution Functions 22
1.5 Measurable Functions and Integration 35
1.6 Basic Integration Theorems 45
1.7 Comparison of Lebesgue and Riemann Integrals 55

2 Further Results in Measure and Integration Theory 60

2.1 Introduction 60
2.2 Radon–Nikodým Theorem and Related Results 64
2.3 Applications to Real Analysis 72
2.4 L^p Spaces 84
2.5 Convergence of Sequences of Measurable Functions 96
2.6 Product Measures and Fubini's Theorem 101
2.7 Measures on Infinite Product Spaces 113
2.8 Weak Convergence of Measures 121
2.9 References 125

3 Introduction to Functional Analysis 127

3.1 Introduction 127
3.2 Basic Properties of Hilbert Spaces 130
3.3 Linear Operators on Normed Linear Spaces 141
3.4 Duality Theorems in Functional Analysis 152
3.5 References 165

4 Basic Concepts of Probability 166

4.1 Introduction 166
4.2 Some Probability Spaces 167
4.3 Independence 167
4.4 Bernoulli Trials 170
4.5 Conditional Probability 171
4.6 Random Variables 173
4.7 Random Vectors 175
4.8 Independent Random Variables 176
4.9 Some Examples from Basic Probability 181
4.10 Expectation 183
4.11 Infinite Sequences of Random Variables 196
4.12 References 200

5 Conditional Probability and Expectation 201

5.1 Introduction 201
5.2 Approximations 202
5.3 The General Concept of Conditional Probability and Expectation 204
5.4 Conditional Expectation Given a σ-Field 214
5.5 Properties of Conditional Expectation 221
5.6 Regular Conditional Probabilities 225

6 Strong Laws of Large Numbers and Martingale Theory 235

6.1 Introduction 235
6.2 Convergence Theorems 236
6.3 Martingales 243
6.4 Martingale Convergence Theorems 250
6.5 Uniform Integrability 262
6.6 Uniform Integrability and Martingale Theory 266
6.7 Optional Sampling Theorems 270
6.8 Applications of Martingale Theory 277
6.9 Applications to Markov Chains 285
6.10 References 294

7 The Central Limit Theorem 296

7.1 Introduction 296
7.2 The Fundamental Weak Convergence Theorem 296
7.3 Characterization of Normal Distributions 307

7.4 Stable Distributions 317
7.5 Infinitely Divisible Distributions 320
7.6 Uniform Convergence in the Central Limit Theorem 329
7.7 Less Smoothed Convergence and Other Convergence Theorems 332
7.8 The Multivariate Central Limit Theorem 336
7.9 References 345

8 Ergodic Theory 345

8.1 Introduction 345
8.2 Ergodicity and Mixing 350
8.3 The Poincaré Ergodic Theorem 356
8.4 Applications to Markov Chains 368
8.5 The Subadditive Ergodic Theorem 377
8.6 Entropy of a Transformation 386
8.7 Bernoulli Shifts 402
8.8 References 407

9 Brownian Motion and Stochastic Integrals 399

9.1 Stochastic Processes 399
9.2 Brownian Motion 401
9.3 Nondifferentiability and Quadratic Variation of Paths 416
9.4 Law of the Iterated Logarithm 430
9.5 The Markov Property 435
9.6 Martingales 440
9.7 Stochastic Integrals 456
9.8 Itô's Differentiation Formula 482
9.9 References 497

Appendices 538

1 The Symmetric Random Walk in $\mathbb{R}^k$ 538
2. Semicontinuous Functions 541
3. Completion of the Proof of Theorem 7.3.2 545
4 Proof of the Cramér-Wold Device Theorem 7.9.1 547
5 The Skorokhod Representation 549

Bibliography 553

Solutions to Problems 566

Index 597

Preface

This text grew out of lecture notes for courses in measure and probability theory taught to graduate students in mathematics and statistics and frequently also to graduate students in electrical engineering. In writing the book, I was motivated by the following objectives:

1. To present the basic results concerning independence, laws of large numbers, characteristic functions, and central limit theorems as quickly as possible and in the most elementary setting.
2. To present measure theory only after some of the basic results of probability theory have been obtained, so that measure theory can be motivated by specific problems.
3. To present enough measure theory to understand modern probability theory.
4. To present some of the interplay between measure theory and probability theory.
5. To present a large number of examples of the use of the theorems.
6. To present a few applications of probability theory to information theory and ergodic theory.

The first edition of this book was published by Addison-Wesley in 1968. The second edition was published by Harcourt Brace Jovanovich in 1972. The third edition was published by Harcourt Brace Jovanovich in 1988. The fourth edition was published by Academic Press in 1995. The present edition is a corrected version of the fourth edition.

Chapters 1 and 2 develop the fundamentals of measure and integration theory, including the theorems of Radon–Nikodym and Fubini, which are essential for understanding the foundations of probability. The Kolmogorov extension theorem, which relates consistent families of finite-dimensional distributions to measures on infinite-dimensional spaces, is proved in Section 2.5. Chapter 3 introduces the basic concepts of probability theory, including independence, conditional probability, and expectation. The first Borel–Cantelli lemma and its application to problems involving almost sure convergence are presented early. The strong law of large numbers is proved in Section 3.4 under the assumption of finite fourth moments. Necessary and sufficient conditions for the strong law are given in Section 5.4. Chapter 4 discusses weak convergence of probability measures, including characteristic functions, the central limit theorem, and infinitely divisible laws. Chapter 5 treats the central limit problem using martingale methods. The martingale convergence theorem is proved, and applications to the law of large numbers, to stopping times, and to the central limit theorem are discussed. Chapter 6 contains an introduction to ergodic theory, including the Birkhoff ergodic theorem and the Shannon–McMillan–Breiman theorem. Chapter 7 discusses the Kolmogorov consistency theorem for infinite product spaces, the Hewitt–Savage zero–one law, and de Finetti's theorem for exchangeable sequences of random variables. Chapter 8 discusses conditional expectation for general random variables, regular conditional probabilities, and conditional independence. Chapter 9 discusses the Skorokhod representation theorem and the martingale characterization of Brownian motion. Chapter 10 discusses the Wiener measure and Donsker's theorem. Chapter 11 discusses the Poisson random measure and its use in modeling jump discontinuities in stochastic processes. Chapter 12 discusses the Ornstein–Uhlenbeck process and the Feynman–Kac formula.

Chapters 8 and 9 have been added in the second edition and should be of interest to the entire prospective audience: mathematicians, statisticians and engineers. Chapter 8 introduces ergodic theory, which is developed for simple Markov chains, with information theory as a guiding motivation. The Shannon McMillan theorem is proved and the isomorphism problem for Bernoulli shifts is discussed. Chapter 9 treats the one dimensional Ising model in statistical mechanics, first in nonrigorous fashion via Peierls arguments and then using the Perron Frobenius theorem.

To make room for the new material, the appendix on general topology and the old Chapter 4 on the interplay between measure theory and topology have been removed. Gone with the section on topological vector spaces is the Hahn Banach theorem, but the more useful version for Hilbert spaces is retained. Also gone is the material on Markov chains, which is now covered in greater depth in Chapter 8. All the necessary background appears in Karl Petersen's *Ergodic Theory* (Cambridge Studies in Advanced Mathematics 2, Cambridge University Press, 1983). The few exercises that require additional background are marked with an asterisk.

It is theoretically possible to read the text without any prior exposure to probability, picking up the necessary equipment in Chapter 4. But we expect that in practice almost all readers will have taken a standard undergraduate course in probability theory. We believe that previous acquaintance with discrete state Markov chains, and random walks are helpful but not essential. A probability course without measure theory will find this unappealing and unbalanced. The summary we have in mind is given in Appendix 1.

Problems are given at the end of each section. Fairly detailed solutions are given to many problems, and instructors may obtain solutions to those problems in Chapters 1–8 not worked out in the text by writing to the publishers.

Catherine Douvas made many Chapter 9, and offered valuable advice and criticism for the other chapters. Mel Gartland kindly allowed us to use material from *Topics in Stochastic Processes* by Ash and Gardner in the text in Chapter 9. We appreciate the encouragement and support provided by the staff at Harcourt/Academic Press.

Robert B. Ash
Catherine Doléans-Dade
Urbana, Illinois, 1999

Summary of Notation

We indicate here the notational conventions to be used throughout the book. The numbering system is standard, for example, 2.3.4 means Chapter 2, Section 3, Part 4. In the appendices, the letter A is used, thus A2.3 means Part 3 of Appendix 2.

The symbol □ is used to mark the end of a proof.

1 Sets

If A and B are subsets of a set Ω, then $A \cup B$ will denote the union of A and B, and $A \cap B$ the intersection of A and B. The union and intersection of a family of sets A_α are denoted by $\bigcup_\alpha A_\alpha$ and $\bigcap_\alpha A_\alpha$. The complement of A relative to Ω is denoted by A^c.

The statement "A is a subset of B" is denoted by $A \subset B$; the inclusion need not be proper. Thus we have $A \subset A$ for any set A. We also write $B \supset A$ or $A \subseteq B$ if B is the result of adjoining to A zero or more elements of B.

The notation $A - B$ will always mean, unless otherwise specified, the set of points that belong to A but not to B. It is referred to as the difference between A and B; a proper difference is a set $A - B$ where $B \subset A$.

The symmetric difference between A and B is by definition the union of $A - B$ and $B - A$; it is denoted by $A \triangle B$.

If $A_1 \subset A_2 \subset \cdots$ and $\bigcup_1^\infty A_n = A$, we say that the A_n form an increasing sequence of sets (increasing to A) and write $A_n \uparrow A$. Similarly if $A_1 \supset A_2 \supset \cdots$ and $\bigcap_1^\infty A_n = A$, we say that the A_n form a decreasing sequence of sets (decreasing to A) and write $A_n \downarrow A$.

The word "finite" will always imply a unique measure, and the word "countable" a countable ordering given. Thus if Ω and $\mathcal{U}$ are collections of sets, "$\mathcal{U}$ includes Ω" means that $\mathcal{U} \supset \Omega$. Equivalently, we may say that $\mathcal{U}$ contains all sets in Ω. In this words each set in Ω is always a member of $\mathcal{U}$.

A countable set is one that is either finite or countably infinite.

The empty set $\emptyset$ is the set with no members. The sets A_α, $\alpha \in I$, are disjoint if $A_\alpha \cap A_\beta = \emptyset$ for all $\alpha \neq \beta$.

2 Real Numbers

The set of real numbers will be denoted by $\mathbb{R}$ and n-dimensional Euclidean space by $\mathbb{R}^n$. If $a, b \in \mathbb{R}$, the interval $[a, b]$ is defined as the set $\{x \in \mathbb{R} : a \leq x \leq b\}$ and $(a, b]$ as $\{x \in \mathbb{R} : a < x \leq b\}$. Other types of intervals are defined similarly. If $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ are points in $\mathbb{R}^n$, $a \leq b$ means $a_i \leq b_i$ for $i = 1, \ldots, n$. The interval $[a, b]$ is defined as $\{x \in \mathbb{R}^n : a_i \leq x_i \leq b_i, i = 1, \ldots, n\}$, and other types of intervals are defined similarly.

The set of extended real numbers is the two-point compactification of $\mathbb{R}$ which is $[-\infty, +\infty]$, denoted by $\overline{\mathbb{R}}$; the set of n-tuples $(x_1, \ldots, x_n)$, with each $x_i \in \overline{\mathbb{R}}$ is denoted by $\overline{\mathbb{R}}^n$. We admit the following rules of arithmetic on $\overline{\mathbb{R}}$:

$$x + \infty = \infty \quad x - \infty = -\infty, \qquad x - \infty = -\infty \quad x + \infty = \infty, \qquad x \in \mathbb{R},$$

$$\infty + \infty = \infty, \qquad \infty \cdot \infty = \infty \qquad (\infty \cdot x \text{ not defined})$$

$$\lambda \cdot \infty = \infty \cdot \lambda = \begin{cases} \infty & \text{if } \lambda > 0, \\ -\infty & \text{if } \lambda < 0, \end{cases} \qquad \lambda > 0,$$

$$\frac{x}{\infty} = \frac{x}{-\infty} = 0 \quad x \in \mathbb{R}, \qquad \left(\frac{\infty}{\infty} \text{ not defined}\right).$$

$$0 \cdot \infty = \infty \cdot 0 = 0$$

The rules are inconsistent when developing the properties of the abstract Lebesgue integral, but it could be emphasized that $\mathbb{R}$ is not a field under these operations.

Unless otherwise specified, measurable means (Borel) measurable and nonnegative means greater than or equal to zero.

The real or complex zero sequence is denoted by 0, and the set of n-tuples of complex numbers by $\mathbb{C}^n$.

3 Functions

If f is a function from Ω_1 to Ω_2 (written as $f : \Omega_1 \to \Omega_2$) and $B \subset \Omega_2$ the preimage of B under f is given by $f^{-1}(B) = \{x \in \Omega_1 : f(x) \in B\}$. It follows from this definition that $f^{-1}\left(\bigcup_\alpha B_\alpha\right) = \bigcup_\alpha f^{-1}(B_\alpha)$, $f^{-1}\left(\bigcap_\alpha B_\alpha\right) = \bigcap_\alpha f^{-1}(B_\alpha)$, $f^{-1}(B_1 - B_2) = f^{-1}(B_1) - f^{-1}(B_2)$; hence $f^{-1}(B^c) = [f^{-1}(B)]^c$. If $\mathcal{C}$ is a class of subsets of Ω_2, $f^{-1}(\mathcal{C})$ means the collection of sets $f^{-1}(B)$, $B \in \mathcal{C}$.

If $f : \mathbb{R} \to \mathbb{R}$, f is increasing if $x \leq y$ implies $f(x) \leq f(y)$; f is decreasing if $x \leq y$ implies $f(x) \geq f(y)$. Thus "increasing" and "decreasing" do not have the strict connotation. If $f_n : \Omega \to \mathbb{R}$, $n = 1, 2, \ldots$, the f_n are said to be increasing if $f_n(x) \leq f_{n+1}(x)$ for all n and x; decreasing sequences are defined similarly.

If f and g are functions from Ω to $\mathbb{R}$, statements such as $f \le g$ are always interpreted as holding pointwise, that is, $f(\omega) \le g(\omega)$ for all $\omega \in \Omega$. Similarly if $f_\alpha : \Omega \to \mathbb{R}$ for each $\alpha \in I$, $\sup_\alpha f_\alpha$ is the function whose value at ω is $\sup\{f_\alpha(\omega) : \alpha \in I\}$.

If $f_1, f_2, \ldots$ is an increasing sequence of functions with limit f (that is, $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ for each ω), we write $f_n \uparrow f$. Similarly, $f_n \downarrow f$ is used for a decreasing sequence.

Sometimes we use such as $\{\omega \in \Omega : f(\omega) \le g(\omega)\}$ briefly written as $[f \le g]$; similarly the preimage $\{\omega \in \Omega : f(\omega) = h\}$ is written as $[f = h]$.

If $A \subseteq \Omega$, the *indicator* or *characteristic function* of A, denoted by 1_A, is defined by $1_A(\omega) = 1$ if $\omega \in A$ and by $1_A(\omega) = 0$ if $\omega \notin A$. The phrase "characteristic function" is often used in the literature, but we shall not adopt this usage here.

If $f(x, y)$ is a function of two variables x and y, the symbol $f(x, \cdot)$ is used for the mapping $y \to f(x, y)$ with x fixed.

The composition of two functions $X : \Omega \to \Omega'$ and $g : \Omega' \to \Omega''$ is denoted by $g \circ X$ or $g(X)$.

If $f : \Omega \to \overline{\mathbb{R}}$, the *positive* and *negative parts* of f are defined by $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$, so that

$$f^+(\omega) = \begin{cases} f(\omega) & \text{if } f(\omega) > 0, \\ 0 & \text{if } f(\omega) \le 0, \end{cases}$$

$$f^-(\omega) = \begin{cases} -f(\omega) & \text{if } f(\omega) < 0, \\ 0 & \text{if } f(\omega) \ge 0. \end{cases}$$

4 Topology

A *topological space* is a set Ω on which is defined a collection $\mathcal{T}$ of subsets of Ω (called the *open sets*), satisfying $\emptyset \in \mathcal{T}$, $\Omega \in \mathcal{T}$, $G_1 \cap G_2 \in \mathcal{T}$ if $G_1, G_2 \in \mathcal{T}$, and $\bigcup_\alpha G_\alpha \in \mathcal{T}$ if $G_\alpha \in \mathcal{T}$ for each α. The collection $\mathcal{T}$ can be chosen in many ways, but usually we are interested only in those $\mathcal{T}$ that satisfy the remaining properties, and it is called a *topology*. The terminology is often used for the pair $(\Omega, \mathcal{T})$.

A *ball* (or *open ball*) in a metric or pseudometric space is a set of the form $B(x, r) = \{y \in \Omega : d(x, y) < r\}$, where x, the *center* of the ball, is a point of Ω and r, the *radius*, is a positive real number. A *closed ball* is a set of the form $\bar{B}(x, r) = \{y \in \Omega : d(x, y) \le r\}$.

Sequences in $\mathbb{R}$ are denoted by $\{x_n, n = 1, 2, \ldots\}$. The term "lower semicontinuous" is abbreviated LSC, and "upper semicontinuous" is abbreviated USC.

An knowledge of general topology (beyond metric spaces) is assumed, and the few comments that refer to general topological spaces can safely be ignored.

5 Vector Spaces

The terms "vector space" and "linear space" are synonymous. All vector spaces are over the real or complex field, and the complex field is assumed unless the term "real vector space" is used.

A *finite basis* for a vector space V is a maximal linearly independent subset of V, i.e., a linearly independent set B such that if $x_1, \ldots, x_n \in B$ and $y \in V$, then $y, x_1, \ldots, x_n$ are not linearly independent. Equivalently, a finite basis is a linearly independent subset B of V with the property that each $x \in V$ is a finite linear combination of elements of B. An infinite-dimensional Banach space (Chapter 5) has different concepts.

The terms "subspace" and "linear manifold" are synonymous: each refers to a subset M of a vector space V that is itself a vector space under the operations of addition and scalar multiplication in V. If there is a finite set B of vectors in M, then M is called a *closed subspace*.

If S is an arbitrary subset of V, the *linear manifold generated by* S, denoted by $L(S)$, is the smallest linear manifold containing all elements of S, that is, the collection of finite linear combinations of elements of S. Analogously, if M is a linear manifold in V, the *space spanned by* M, denoted by $S(M)$, is the smallest closed subspace containing all elements of M. Equivalently, $S(M)$ is the closure of $L(M)$.

6 Zorn's Lemma

A *partial ordering* on a set S is a relation "$\leq$" that is

(1) reflexive: $a \leq a$,
(2) antisymmetric: if $a \leq b$ and $b \leq a$, then $a = b$, and
(3) transitive: if $a \leq b$ and $b \leq c$, then $a \leq c$

(all elements a, b, c belong to S).

If $C \subseteq S$, C is said to be *totally ordered* iff for all $a, b \in C$, either $a \leq b$ or $b \leq a$. A totally ordered subset of S is also called a *chain* in S.

The form of Zorn's lemma that will be used in the text is as follows.

Let S be a set with a partial ordering "$\leq$." Assume that every chain C in S has an upper bound in S; in other words, there is an element $a \in S$ such that $x \leq a$ for all $x \in C$. Then S has a maximal element, that is, an element $m \in S$ such that for each $x \in S$ it is not possible to have $m \leq x$ and $m \neq x$.

Zorn's lemma is usually an axiom of set theory, equivalent to the axiom of choice.

CHAPTER

1

FUNDAMENTALS OF MEASURE AND INTEGRATION THEORY

In this chapter we give a self-contained presentation of the basic concepts of the theory of measure and integration. The principles learned here and in Chapter 2 will serve as background for the study of probability as well as harmonic analysis, linear space theory, and other areas of mathematics.

1.1 [illegible]

It will be convenient to start with a little practice in the algebra of sets. This will serve as a refresher and also as a way of collecting a few results that will often be useful.

Let $A_1, A_2, \ldots$ be subsets of a set X. If $A_1 \subset A_2 \subset \cdots$ and $\bigcup_{n=1}^{\infty} A_n = A$, we say that the A_n form an increasing sequence of sets with limit A, or that the A_n increase to A, and write $A_n \uparrow A$. If $A_1 \supset A_2 \supset \cdots$ and $\bigcap_{n=1}^{\infty} A_n = A$, we say that the A_n form a decreasing sequence of sets with limit A, or that the A_n decrease to A; we write $A_n \downarrow A$.

The De Morgan laws remind us that $\left(\bigcup_n A_n\right)^c = \bigcap_n A_n^c$ and $\left(\bigcap_n A_n\right)^c = \bigcup_n A_n^c$. It is [illegible]

(1) If $A_n \uparrow A$, then $A_n^c \downarrow A^c$; if $A_n \downarrow A$, then $A_n^c \uparrow A^c$.

It is sometimes useful to write a union of sets as a disjoint union. This may be done as follows.

Let $A_1, A_2, \ldots$ be subsets of X. For each n we have

$$(2) \quad \bigcup_{j=1}^{n} A_j = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_1 \cup A_2)) \cup \cdots \cup (A_n \setminus (A_1 \cup \cdots \cup A_{n-1}))$$

and

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_1 \cup A_2)) \cup \cdots$$

The statements

$$(3) \quad \bigcup_{j=1}^{n} A_j = A_1 \cup (A_2 \setminus A_1) \cup \cdots \cup (A_n \setminus A_{n-1})$$

and

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup (A_2 \setminus A_1) \cup \cdots$$

are also correct. In (2) and (3), the sets on the right are disjoint. If the A_j form an increasing sequence, the formulas become

(1.1) $\bigcap_{n\in\mathbb{N}} A_n = \{x : (\forall n \in \mathbb{N})(x \in A_n)\} = \{x : x \in A_1, x \in A_2, \ldots\}$

and

(1.2) $\bigcup_{n\in\emptyset} A_n = \bigcap_{n\in\emptyset} A_n = \{x : (\exists n \in \emptyset)(x \in A_n)\}$

(since $\emptyset$ is the empty set).

The results (1.1)–(1.2) are proved using only the definitions of union, intersection, and complementation, see Problem 1.

The following set operations will be of particular interest. If $A = \{A_n\}_{n\in\mathbb{N}}$ is a sequence of sets, we define

(1.3) $\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$.

Thus $x \in \limsup_n A_n$ iff for every n, $x \in A_k$ for some $k \geq n$, in other words

(1.4) $x \in \limsup_n A_n$ iff $x \in A_n$ for infinitely many n.

Also define

(1.5) $\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$.

Thus $x \in \liminf_n A_n$ iff for some n, $x \in A_k$ for all $k \geq n$, in other words

(1.6) $x \in \liminf_n A_n$ iff $x \in A_n$ eventually, that is, for all but finitely many n.

We shall call $\limsup_n A_n$ the upper limit of the sequence of sets A_n, and $\liminf_n A_n$ the lower limit. This terminology is, of course, suggested by the analogous concepts for sequences of real numbers

$$\limsup_n a_n = \inf_n \sup_{k\geq n} a_k$$

$$\liminf_n a_n = \sup_n \inf_{k\geq n} a_k$$

See Problem 4 for a further development of this analogy.

The following rules may be verified (Problem 5):

(1.7) $\left(\limsup_n A_n\right)^c = \liminf_n A_n^c$

(1.8) $\left(\liminf_n A_n\right)^c = \limsup_n A_n^c$

(1.9) $\limsup_n A_n = \limsup_n A_{n+1}$

(1.10) If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and $\bigcup_{n=1}^{\infty} A_n = A$, then $\liminf_n A_n = \limsup_n A_n = A$.

In general, if $\limsup_n A_n = \liminf_n A_n = A$, then A is said to be the *limit* of the sequence $A_1, A_2, \ldots$, and we write $A = \lim_n A_n$.

Problems

1 Verify the formulas (1.1)–(1.2).

2 Define a sequence of real numbers as follows: Let $a_n = (-1)^n$ if n is odd, and $a_n = (-1, 1)$ if n is even. Find $\limsup_n a_n$ and $\liminf_n a_n$.

3 Let $S = \mathbb{R}^2$, A_1 the interior of the circle with center at $(-1, 0)$ and radius 1, and A_2 the interior of the circle with center at $(1, 0)$ and radius 1. Find $\limsup_n A_n$ and $\liminf_n A_n$.

4. Let $\{x_n\}$ be a sequence of real numbers, and let $A_n = \{x_n, x_{n+1}, \ldots\}$. What is the connection between $\limsup_{n\to\infty} x_n$ and $\inf_n \sup_{k\geq n} x_k$? Similarly for $\liminf_{n\to\infty} x_n$ and $\sup_n \inf_{k\geq n} x_k$?
5. Establish formulas (1.1) and (1.2).
6. Let $A_n = [0, \frac{1}{n}]$ and $B_n = (\frac{1}{n}, 2]$ be subsets of $\mathbb{R}$, and let $C_n = A_n \cup B_n$ if n is odd, $C_n = \emptyset$ if n is even. Find $\limsup_n C_n$ and $\liminf_n C_n$.

1.2 Fields, σ-Fields, and Measures

Examples such as volume, area, and probability suggest the existence of a class of sets $\mathcal{A}$ and a nonnegative set function μ defined on $\mathcal{A}$ with certain properties. A measure is a set function that is, an assignment of a number, $\mu(A)$, to each set A in a certain class. Some structure must be imposed on the class of sets on which μ is defined, and probabilistic considerations alone do a good job of motivating the type of structure required. If Ω is a set whose points correspond to the possible outcomes of a random experiment, certain subsets of Ω correspond to "events" and we regard a probability as a function defined on the events. Intuitively, if A is an event, the question "does ω belong to A?" has a definite yes or no answer after the experiment is performed (and the outcome ω is known). Now if A and B are events, the question "does $\omega \in A \cup B$?" can certainly be answered: the question "is $\omega \in A$?" and, for each $n = 1, 2, \ldots$, we can decide whether or not ω belongs to A_n, then we can determine whether or not ω belongs to $\bigcup_1^\infty A_n$. (and similarly for $\bigcap_1^\infty A_n$). Thus it is natural to require that the class of events be closed under complementation, finite unions, and more generally countable unions and intersections. Furthermore, since the answer to the question "is $\omega \in \Omega$?" is always "yes," the entire space Ω should be an event. Closure under countable unions and intersections is difficult to justify physically, and perhaps the most convincing reason for requiring it is that a richer mathematical theory is obtained. Fortunately, we are able to assert that the limit of a sequence of events is an event; see 1.3.1.

1.2.1 *Definition.* Let $\mathcal{F}$ be a collection of subsets of a set Ω. Then $\mathcal{F}$ is called a *field* (or *algebra*) if it contains Ω and is closed under complementation and finite unions, that is,

(a) $\Omega \in \mathcal{F}$.
(b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(c) If $A_1, \ldots, A_n \in \mathcal{F}$, then $\bigcup_1^n A_i \in \mathcal{F}$.

It follows that $\mathcal{F}$ is closed under finite intersections. For if $A_1, \ldots, A_n \in \mathcal{F}$, then

$$\bigcap_1^n A_i = \left(\bigcup_1^n A_i^c\right)^c \in \mathcal{F}$$

It is natural to strengthen (c) to countable unions, that is,

(c) If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

$\mathcal{F}$ is called a *field* over X or an *algebra* over X if it is merely required to be closed under finite unions. $\mathcal{F}$ is also closed under countable intersection.

If $\mathcal{F}$ is a field and countable collection of sets in $\mathcal{F}$ can be expressed as the limit of an increasing sequence of sets in $\mathcal{F}$, and conversely. To see this note that if $A = \bigcup_{n=1}^{\infty} A_n$ then $A = \lim_{n\to\infty} A_n$ if monotonically. If $A_n \uparrow A$, then $A = \bigcup_{n=1}^{\infty} A_n$. This shows that a field is a field that is closed under limits of increasing sequences.

1.2.2 *Examples.* The largest σ-field of subsets of a fixed set Ω is the collection of all subsets of Ω. The smallest σ-field consists of the two sets $\emptyset$ and Ω.

Let Ω be a nondenumerable pointset of Ω, and let $\mathcal{F} = \{\emptyset, \Omega, A, A^c\}$. Then $\mathcal{F}$ is the smallest σ-field containing A. For if $\mathcal{S}$ is a σ-field and $A \in \mathcal{S}$, then by definition of σ-field, $\Omega \in \mathcal{S}$, and A^c belongs to $\mathcal{S}$ because $A^c \in \mathcal{S}$. But $\mathcal{F}$ is a σ-field, for all the four complements or unions of sets in $\mathcal{F}$, we correctly obtain sets in $\mathcal{F}$. Thus $\mathcal{F}$ is the desired one that is included in any σ-field containing A, and the result follows.

If $A_1, \ldots, A_n$ are arbitrary subsets of Ω, the smallest σ-field containing $A_1, \ldots, A_n$ may be described explicitly; see Problem 4.

If $\mathcal{C}$ is a class of sets, the smallest σ-field containing the sets in $\mathcal{C}$ will be written as $\sigma(\mathcal{C})$, and sometimes called the *minimal σ-field over $\mathcal{C}$*. We also call $\sigma(\mathcal{C})$ the *σ-field generated by $\mathcal{C}$*; one commonly thinks of $\mathcal{C}$ as being a relatively uncomplicated collection.

Let Ω be the set $\mathbb{R}$ of real numbers. Let $\mathcal{C}$ consist of all finite disjoint unions of right-semiclosed intervals. A right-semiclosed interval is a set of the form $(a, b] = \{x : a < x \le b\}$, where $-\infty \le a < b < \infty$; by convention we also count $(a, b]$ as right-semiclosed for $-\infty \le a < b < \infty$. This convention is necessary because $\mathbb{R} = (-\infty, \infty]$ belongs to $\mathcal{C}$, and if $\mathcal{C}$ is to be a field, the complement of $(-\infty, \infty]$ must belong to $\mathcal{C}$. It may be verified that conditions (a), (b), (c) of 1.2.1 hold, and thus $\mathcal{C}$ is a field. Here $\mathcal{C}$ is not a σ-field; for example, $A_n = (0, 1 - 1/n] \in \mathcal{C}$, $n = 1, 2, \ldots$, and $\bigcup_{n=1}^{\infty} A_n = (0, 1) \notin \mathcal{C}$.

If Ω is the set $\mathbb{R} = (-\infty, \infty]$ of extended real numbers, then just as above, the collection $\mathcal{C}$ of finite disjoint unions of right-semiclosed intervals forms a field but not a σ-field. Here, the right-semiclosed intervals are sets of the form $(a, b] = \{x : a < x \le b\}$, $-\infty \le a < b < \infty$, and by convention the sets $(a, \infty] = \{x : a < x \le \infty\}$, $-\infty \le a < \infty$, are also included. (In this case the convention is necessary because $(b, \infty]^c = [-\infty, b]$ must belong to $\mathcal{C}$, and therefore the complement $[-\infty, a]$ also belongs to $\mathcal{C}$.)

There is a type of reasoning that occurs so often in problems involving σ-fields that it deserves to be displayed explicitly. We call this *Dynkin's typical illustration*.

If $\mathcal{C}$ is a class of subsets of Ω and $A \subset \Omega$, we denote by $\mathcal{C} \cap A$ the class $\{B \cap A : B \in \mathcal{C}\}$. If the minimal σ-field over $\mathcal{C}$ is $\sigma(\mathcal{C}) = \mathcal{A}$, let us show that

$$\sigma_A(\mathcal{C} \cap A) = \mathcal{A} \cap A,$$

where $\sigma_A(\mathcal{C} \cap A)$ is the minimal σ-field of subsets of A over $\mathcal{C} \cap A$. In other words, $\mathcal{A} \cap A$ is regarded as the entire space.

Now $\mathcal{C} \cap A \subset \mathcal{A} \cap A$, since $\mathcal{C} \subset \mathcal{A}$, and it is not hard to verify that $\mathcal{A} \cap A$ is a σ-field of subsets of A. Therefore $\sigma_A(\mathcal{C} \cap A) \subset \mathcal{A} \cap A$.

To establish the reverse inclusion we must show that $B \cap A \in \sigma_A(\mathcal{C} \cap A)$ for all $B \in \sigma(\mathcal{C})$. This is not obvious, so we resort to the following basic measuring process, which might be called the good sets principle. Let $\mathcal{F}$ be the class of good sets, that is, let $\mathcal{F}$ consist of those sets $B \subset \Omega$ such that

$$B \cap A \in \sigma_A(\mathcal{C} \cap A).$$

Since $\mathcal{C} \subset \mathcal{F}$ and $\sigma_A(\mathcal{C} \cap A)$ is a σ-field, it follows quickly that $\mathcal{F}$ is a σ-field. Hence $\sigma(\mathcal{C}) \subset \mathcal{F}$ (so that $\sigma(\mathcal{C}) \subset \mathcal{F}$ because $\mathcal{F} \subset \sigma(\mathcal{C})$ is the reverse follows trivially, every set in $\mathcal{C}$ is good and the class of good sets contains a σ-field; consequently, every set in $\sigma(\mathcal{C})$ is good.

One other comment. If $\mathcal{C}$ is closed under finite intersections and $A \in \mathcal{C}$, then $\mathcal{C} \cap A = \{C \cap A : C \in \mathcal{C}\}$ (observe that if $C \in \mathcal{A}$, then $C \cap A = C \cap A$).

1.2.3 Definitions and Comments. A measure on a σ-field $\mathcal{A}$ is a countably additive, nonnegative, extended real-valued function μ on $\mathcal{A}$ such that whenever $A_1, A_2, \ldots$ form a finite or countably infinite collection of disjoint sets in $\mathcal{A}$, we have

$$\mu\left(\bigcup_j A_j\right) = \sum_j \mu(A_j).$$

If $\mu(\Omega) = 1$, μ is called a probability measure.

A measure space is a triple $(\Omega, \mathcal{A}, \mu)$ where Ω is a set, $\mathcal{A}$ is a σ-field of subsets of Ω, and μ is a measure on $\mathcal{A}$. If μ is a probability measure, $(\Omega, \mathcal{A}, \mu)$ is called a probability space.

It will be convenient to have a slight generalization of the notion of a measure on a σ-field. Let $\mathcal{F}$ be a field of subsets of Ω (a set function on $\mathcal{F}$ to map from $\mathcal{F}$ into $\mathbb{R}^*$). We say that μ is countably additive on $\mathcal{F}$ if whenever $A_1, A_2, \ldots$ form a finite or countably infinite collection of disjoint sets in $\mathcal{F}$ whose union also belongs to $\mathcal{F}$ (this will always be the case if $\mathcal{F}$ is a σ-field) we have

$$\mu\left(\bigcup_j A_j\right) = \sum_j \mu(A_j).$$

If this equation holds only for finite collections of disjoint sets in $\mathcal{F}$, μ is said to be finitely additive on $\mathcal{F}$. To avoid the appearance of terms of the form

$+\infty$ or $-\infty$ in the summation, we always assume that $+\infty$ and $-\infty$ cannot both belong to the range of μ.

If μ is merely finitely additive and not σ-additive, μ is called a measure on $\mathcal{A}$, a probability measure if $\mu(\Omega) = 1$.

Note that countable additivity actually implies finite additivity. For if $\mu(A_1) = -\infty$ for all $A_1 \in \mathcal{A}$, then $\mu(A) = -\infty$ for all $A \in \mathcal{A}$, the result is immediate. Therefore assume $\mu(A_1)$ is finite for some $A_1 \in \mathcal{A}$. By considering the sequence $A_1, \emptyset, \emptyset, \ldots$, we find that $\mu(\emptyset) = 0$, and hence additivity is easily established by considering the sequence $A_1, \ldots, A_n, \emptyset, \emptyset, \ldots$ where $A_1, \ldots, A_n$ are disjoint sets in $\mathcal{A}$.

Although the set function given by $\mu(A) = +\infty$ for all $A \in \mathcal{A}$ satisfies the definition of a measure, and similarly $\mu(A) = -\infty$ for all $A \in \mathcal{A}$ defines a countably additive set function, we shall from now on exclude these cases. Therefore by the above discussion, we always have $\mu(\emptyset) = 0$.

1.1.2 *Examples.* Let Ω be any set and let $\mathcal{A}$ consist of all subsets of Ω. Define $\mu(A)$ as the number of points of A. Thus if A has a number $n = 0, 1, 2, \ldots$ then $\mu(A) = n$; if A is an infinite set, $\mu(A) = \infty$. The set function μ is a measure on $\mathcal{A}$, called *counting measure* on Ω.

A closely related measure is defined as follows. Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be a finite or countably infinite set, and let $p_1, p_2, \ldots$ be nonnegative numbers. Take $\mathcal{A}$ to be all subsets of Ω and define

$$\mu(A) = \sum_{\omega_i \in A} p_i$$

Thus if $A = \{\omega_{i_1}, \omega_{i_2}, \ldots\}$, then $\mu(A) = p_{i_1} + p_{i_2} + \cdots$. The set function μ is a measure on $\mathcal{A}$ and $\mu(\{\omega_i\}) = p_i$, $i = 1, 2, \ldots$. A probability measure will be obtained iff $\sum p_i = 1$. If all $p_i = 1$, then μ is counting measure.

Now if A is a subset of $\mathbb{R}$, we say A is an interval if A is of the form (a, b), $[a, b]$, $(a, b]$, or $[a, b)$, where a and b, $a \leq b$, are the usual extended real numbers. Take the length of A to be $b - a$ if A is a bounded set, ∞ if A is unbounded. A set may not have any intuitive idea of its length but we shall see in Section 1.4 that there exists a class $\mathcal{A}$ of subsets of $\mathbb{R}$, including all intervals, and a measure μ on $\mathcal{A}$, Lebesgue measure, determining μ on a large class of sets.

Specifically, μ is characterized on the collection $\mathcal{A}$ of Borel sets of $\mathbb{R}$ (defined by $\mathcal{A}(\mathbb{R})$ and defined as the smallest σ-field of subsets of $\mathbb{R}$ containing all intervals (a, b), $-\infty \leq a \leq b \leq \infty$).

Now that $\mu(A)$ is guaranteed to exist, it may be described additively on a rather elemental level as the intersection of A with fields containing the intervals

$[a,b)$. Also, if a set field contains, say, all open intervals, it must contain all intervals $[a,b)$, and conversely. For

$$[a,b) = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right) \quad \text{and} \quad (a,b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b\right).$$

Thus $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing all open intervals. Similarly we may replace the intervals $(a,b]$ by other classes of intervals, for instance,

all closed intervals,
all intervals $[a,b)$, $a,b \in \mathbb{R}$,
all intervals $(-\infty, a]$, $a \in \mathbb{R}$,
all intervals (a,∞), $a \in \mathbb{R}$,
all intervals $(-\infty, b)$, $b \in \mathbb{R}$,
all intervals $(-\infty, b]$, $b \in \mathbb{R}$.

Since a σ-field that contains all intervals of a given type contains all intervals of any other type, $\mathcal{B}(\mathbb{R})$ may be described as the smallest σ-field that contains the class of all intervals of $\mathbb{R}$. Similarly, $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing all open sets of $\mathbb{R}$. (Recall that an open set is a countable union of open intervals.) Since a set is closed if its complement is closed, $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing all closed sets of $\mathbb{R}$. Finally, if $\mathcal{F}_0$ is the field of finite disjoint unions of right semiclosed intervals (see 1.2.2), then $\mathcal{B}(\mathbb{R})$ is the smallest σ-field containing $\mathcal{F}_0$.

Intuitively, we may think of generating the Borel sets by starting with the intervals and keeping complements, countable unions, and intersections in all possible ways. This idea is made precise in Problem 1.1.

The class of Borel sets of $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is defined as the smallest σ-field of subsets of $\mathbb{R}$ containing all intervals $(a,b]$, $a,b \in \overline{\mathbb{R}}$. This shows immediately concerning the replacement of the right semiclosed intervals by other classes of sets applies equally well in $\mathbb{R}$.

If $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$, $\mathcal{B}(\mathcal{A})$ will denote $\mathcal{A} \cap \mathcal{F}$ $(\mathcal{A} \subset \mathcal{F})$; this coincides with $\{A \cap \mathbb{R} : A \in \mathcal{B}(\overline{\mathbb{R}})\}$ (see 1.2.2).

We may begin to develop some properties of set functions.

1.2.3 *Theorem.* Let μ be a finitely additive set function on the field $\mathcal{F}$.

(a) $\mu(\emptyset) = 0$.
(b) $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$ for all $A, B \in \mathcal{F}$.
(c) If $A, B \in \mathcal{F}$ and $A \subset B$, then $\mu(A) \leq \mu(B)$ and $\mu(B \setminus A) = \mu(B) - \mu(A)$.

(Here $\mu(A) - \mu(B) = \mu(A) + (-\mu(B))$ if $\mu(B)$ is finite, and $\mu(B) - \mu(A)$ if $\mu(A) = \mu(B) = \infty$.)

(c) If μ is nonnegative,

$$\mu\left(\bigcup_{j=1}^{n} A_j\right) \le \sum_{j=1}^{n} \mu(A_j) \quad \text{for all} \quad A_1, \ldots, A_n \in \mathcal{A}.$$

If μ is a measure,

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) \le \sum_{j=1}^{\infty} \mu(A_j)$$

for all $A_1, A_2, \ldots \in \mathcal{A}$ such that $\bigcup_{j=1}^{\infty} A_j \in \mathcal{A}$.

Proof. (a) Pick $A \in \mathcal{A}$ such that $\mu(A)$ is finite; then

$$\mu(A) = \mu(A \cup \emptyset) = \mu(A) + \mu(\emptyset).$$

(b) By finite additivity,

$$\mu(A) = \mu(A \cap B) + \mu(A - B),$$
$$\mu(B) = \mu(A \cap B) + \mu(B - A).$$

Add the above equations to obtain

$$\mu(A) + \mu(B) = \mu(A \cap B) + [\mu(A \cap B) + \mu(A - B) + \mu(B - A)]$$
$$\ge \mu(A \cup B) + \mu(A \cap B).$$

(c) We may write $A = B \cup (A - B)$, a disjoint union, hence $\mu(A) = \mu(B) + \mu(A - B)$, $\mu(A - B) \ge 0$. We have

$$\mu\left(\bigcup_{j=1}^{n} A_j\right) = \mu(A_1) + \mu(A_2 - A_1) + \mu(A_3 - A_1 \cup A_2) + \cdots + \mu(A_n - A_1 \cup \cdots \cup A_{n-1})$$

[see Section 1.1, formula (2)]. The sets on the right are disjoint and

$$\mu(A_1) + \cdots + \mu(A_n) \quad \text{by (b)}.$$

This case in which μ is a measure is handled using identity (2) of Section 1.1. ∎

1.2.4 *Definitions.* A set function μ defined on $\mathcal{A}$ is said to be *positive* if $\mu(A) \ge 0$ for every $A \in \mathcal{A}$. It is *finitely additive* if $\mu(\emptyset) = 0$ and $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $A, B \in \mathcal{A}$ and $A \cap B = \emptyset$. If μ is finitely additive, it is

subideal or impure. Let $\mu(E)$ be finite, for $E = A_1 \cup A_2$, and $\mu(A_j)$ is small. Then $\mu(E)$ is $\mu(A_1)$.

A nonnegative, finitely additive set function over the field $\mathcal{F}$ is said to be σ-finite on $\mathcal{F}$ if X can be written as $\bigcup_{n=1}^{\infty} A_n$, where each A_n belongs to $\mathcal{F}$ and $\mu(A_n) < \infty$ for all n. (By Example (3) of Section 1.1, the A_n may be assumed disjoint.) We shall see that many properties of finite measures can be extended easily to σ-finite measures.

It follows from 1.2.5(c) that a nonnegative, finitely additive set function μ on a field $\mathcal{F}$ is finite if it is bounded; that is, $\sup\{\mu(A) : A \in \mathcal{F}\} < \infty$. This terminology breaks if the nonnegativity assumption is dropped (see Problem 4). It is true, however, that a countably additive set function on a field is finite iff it is bounded; this will be proved in 2.1.3.

Countably additive set functions have a basic continuity property, which we now describe.

1.2.7 **Theorem** Let μ be a countably additive set function on the field $\mathcal{F}$.

(a) If $A_1, A_2, \ldots \in \mathcal{F}$ and $A_n \uparrow A$, then $\mu(A_n) \to \mu(A)$ as $n \to \infty$.

(b) If $A_1, A_2, \ldots \in \mathcal{F}$, $A_n \downarrow A$, and $\mu(A_1)$ is finite, then $\mu(A_n) \to \mu(A)$ as $n \to \infty$. (Since $\mu(A_1)$ is finite, $|\mu(A_n)| < \infty$ for all n since $|\mu(A_n)| = |\mu(A_n) + [\mu(A_1) - \mu(A_1)]|$, then $\mu(A_n) \to \mu(A)$ as $n \to \infty$.)

The same result holds if $\mathcal{F}$ is only assumed to be a field, if we add the hypothesis that the limit set A belongs to $\mathcal{F}$. If $A_n \downarrow \emptyset$ and μ is finite on $\mathcal{F}$, then $A_n \downarrow \emptyset$ implies that $\mu(A_n)$ decreases to a limit as $n \to \infty$, and decreases to 0 in particular, but we cannot identify the limit with $\mu(\emptyset)$.

Proof (a) If $\mu(A_n) = \infty$ for some n, then $\mu(A) = \mu(A_n) = \mu(A - A_n) = \infty$ by 1.2.5(c) and $\mu(A_n) = \infty$. Replacing A_n by A_1 we find that $\mu(A_1) = \infty$ for all $n \geq N$, and $\mu(A_n) = \infty$ initially. In the same way we eliminate the case in which $\mu(A_n) = -\infty$ for some n. Thus we may assume that all $\mu(A_n)$ are finite.

Since the A_n form an increasing sequence, we may use Exercise 15 of Section 1.1:

$$A = A_1 \cup (A_2 - A_1) \cup (A_3 - A_2) \cup \cdots \cup (A_n - A_{n-1}) \cup \cdots.$$

Therefore, by 1.2.5(c),

$$\mu(A) = \mu(A_1) + \mu(A_2 - A_1) + \cdots + \mu(A_n - A_{n-1}) + \cdots$$
$$= \lim_{n} \mu(A_n).$$

(b) If $A_n \downarrow A$, then $A_1 - A_n \uparrow A_1 - A$, hence $\mu(A_1 - A_n) \to \mu(A_1 - A)$ by (a). The result now follows from 1.2.5(c). ∎

We shall frequently encounter situations in which finite additivity is a particularly natural set function property, but countable additivity is more difficult. It is useful to have the result that finite additivity plus continuity implies countable additivity.

1.5F *Theorem.* Let μ be a finitely additive set function on the field $\mathcal{F}$.

(a) Assume that μ is continuous from below on each $A \in \mathcal{F}$. That is, if $A_1, A_2, \ldots \in \mathcal{F}$, $A_1 \subset A_2 \subset \cdots$, and $A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$, then $\mu(A_n) \to \mu(A)$. Then μ is countably additive on $\mathcal{F}$.

(b) Assume that μ is continuous from above at the empty set. That is, if $A_1, A_2, \ldots \in \mathcal{F}$, $A_1 \supset A_2 \supset \cdots$, and $\bigcap_{n=1}^{\infty} A_n = \emptyset$, then $\mu(A_n) \to 0$. It follows that μ is countably additive on $\mathcal{F}$.

Proof. (a) Let $A_1, A_2, \ldots$ be disjoint sets in $\mathcal{F}$ whose union A belongs to $\mathcal{F}$. If $B_n = \bigcup_{i=1}^{n} A_i$, then $B_n \in \mathcal{F}$, hence $\mu(B_n) \to \mu(A)$ by hypothesis. But $\mu(B_n) = \sum_{i=1}^{n} \mu(A_i)$ by finite additivity, hence $\mu(A) = \lim_{n \to \infty} \sum_{i=1}^{n} \mu(A_i)$, the desired result.

(b) Let $A_1, A_2, \ldots$ be disjoint sets in $\mathcal{F}$ whose union A belongs to $\mathcal{F}$, and let $B_n = \bigcup_{i=n+1}^{\infty} A_i$. By finite additivity, $\mu(A) = \mu(B_n) + \mu(A - B_n)$, but $A - B_n \downarrow \emptyset$, so by hypothesis, $\mu(A - B_n) \to 0$. Thus $\mu(B_n) \to \mu(A)$, and the remainder follows as in (a). ∎

If μ_1 and μ_2 are measures on a field $\mathcal{F}$, then $\mu = \mu_1 - \mu_2$ is countably additive on $\mathcal{F}$, assuming either μ_1 or μ_2 is finite valued. We shall see later (in 3.1F) that any countably additive set function on a field can be expressed as the difference of two measures.

For examples of finitely additive set functions that are not countably additive, see Problems 1.3 and 1.4.

Problems

1. Let Ω be a countably infinite set, and let $\mathcal{F}$ consist of all subsets of Ω. Define $\mu(A) = 0$ if A is finite, and $\mu(A) = \infty$ if A is infinite.
   (a) Show that μ is finitely additive but not countably additive.
   (b) Show that Ω is the union of an increasing sequence of sets A_n with $\mu(A_n) = 0$ for all n, but $\mu(\Omega) = \infty$.
2. Let μ be a nonnegative measure on $\mathcal{F}$, where Ω is an arbitrary set. Show that μ is a semimetric on $\mathcal{F}$ if $d(A, B) = \mu(A \triangle B)$, $A, B \in \mathcal{F}$.
3. Let Ω be a countably infinite set, and let $\mathcal{F}$ be the field consisting of all finite subsets of Ω and their complements. If A is finite, set $\mu(A) = 0$, and if A^c is finite, set $\mu(A) = 1$.
   (a) Show that μ is finitely additive but not countably additive on $\mathcal{F}$.

(b) Show that $\mathcal{C}$ is the π-field of an increasing sequence of sets $A_1 \subset A_2 \subset \cdots$ with $m(A_n) < \infty$ for all n, but $\mu(\Omega) = \infty$.

4. Let $\mathcal{F}$ be the set of finite disjoint unions of right-semiclosed intervals of $\mathbb{R}$, and define the set function μ on $\mathcal{F}$ as follows:

$$\mu(\emptyset) = 0,$$

$$\mu(a, b] = b - a, \qquad a, b \in \mathbb{R},\ a \le b,$$

$$\mu(a, \infty) = \infty, \qquad a \in \mathbb{R},$$

$$\mu(\mathbb{R}) = \infty,$$

$$\mu\left(\bigcup_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} \mu(I_j)$$

if $I_1, \ldots, I_n$ are disjoint right semiclosed intervals.

(a) Show that μ is finitely additive but not countably additive on $\mathcal{F}$.
(b) Show that μ is σ-finite but not extendable to $\mathcal{P}$.

5. Let μ be a nonnegative, finitely additive set function on the field $\mathcal{F}$. If $A, A_1, A_2, \ldots$ are disjoint sets in $\mathcal{F}$ and $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$, show that

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \ge \sum_{n=1}^{\infty} \mu(A_n).$$

6. Let $T: \Omega \to \Omega$, and let $\mathcal{C}$ be a class of subsets of Ω. Show that

$$T^{-1}(\sigma(\mathcal{C})) = \sigma(T^{-1}(\mathcal{C})),$$

where $T^{-1}(\mathcal{C}) = \{T^{-1}(A) : A \in \mathcal{C}\}$. (Use the good sets principle.)

7. If A is a Borel subset of $\mathbb{R}$, show that the smallest σ-field of subsets of A containing the sets open in A (in the relative topology inherited from $\mathbb{R}$) is $\{B \cap A : B \in \mathcal{B}(\mathbb{R})\}$.

8. Let $A_1, \ldots, A_n$ be arbitrary subsets of a set Ω. To describe explicitly the smallest σ-field $\mathcal{F}$ containing $A_1, \ldots, A_n$. How many sets are there in $\mathcal{F}$? (Give an upper bound that is attainable under certain conditions.) List all the sets in $\mathcal{F}$ when $n = 2$.

9. (a) Let $\mathcal{C}$ be an arbitrary class of subsets of Ω, and let $\mathcal{C}'$ be the collection of all finite unions $\bigcup_{i=1}^{n} A_i$, $n = 1, 2, \ldots$, where each A_i is a finite intersection $\bigcap_{j=1}^{m_i} B_{ij}$ with B_{ij} or its complement a set in $\mathcal{C}$. Show that $\mathcal{C}'$ is the minimal field that contains $\mathcal{C}$.

(b) Show that the minimal field can also be described as the collection of all finite disjoint unions $\bigcup_{i=1}^{n} A_i$, where the A_i are as above.

(a) If $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are fields of subsets of Ω, show that the smallest field including $\mathcal{F}_1, \ldots, \mathcal{F}_n$ consists of all finite disjoint unions of sets $A_1 \cap \cdots \cap A_n$ with $A_i \in \mathcal{F}_i$ or $A_i^c \in \mathcal{F}_i$, $i = 1, \ldots, n$.

(b) Let μ be a finite measure on the smallest σ-field containing $\mathcal{F}_1, \ldots, \mathcal{F}_n$, and let $\pi = \Pi_{i=1}^n \mathcal{F}_i$ (see Section 1.1). Show that $\mu(\mathcal{A}) = \lim_{\pi \ni \mathcal{B} \uparrow \mathcal{A}} \mu(\mathcal{B})$.

1.7 Let $\mathcal{C}$ be any class of subsets of Ω, with $\emptyset, \Omega \in \mathcal{C}$. Define $\mathcal{C}_0 = \mathcal{C}$, and for any ordinal $\alpha > 0$ define inductively

$$\mathcal{C}_\alpha = \left\{ \bigcup_{\beta < \alpha} \left(A_\beta \setminus B_\beta \right) : A_\beta, B_\beta \in \mathcal{C}_\beta \right\},$$

where $\mathbb{N}$ denotes the class of all countable ordinal numbers. Show that $\mathcal{C}_\alpha = \mathcal{C}_{\omega_1}$ for all $\alpha \geq \omega_1$.

Let $\mathcal{F} = \bigcup_{\alpha < \omega_1} \mathcal{C}_\alpha$. Where ω_1 is the first uncountable ordinal, and let $\mathcal{F}$ be the minimal σ-field over $\mathcal{C}$. Since each $\mathcal{C}_\alpha \subseteq \mathcal{F}$ we have $\mathcal{F} \subseteq \mathcal{F}$. Also, the $\mathcal{C}_\alpha$ increase with α, and $\mathcal{C}_\alpha \subseteq \mathcal{C}_{\alpha+1}$ for all α.

(a) Show that $\mathcal{F}$ is a σ-field (hence $\mathcal{F} = \mathcal{F}$ by minimality of $\mathcal{F}$).

(b) If the cardinality of $\mathcal{C}$ is at most $\aleph_1$, the cardinality of the reals, show that $\mathcal{F} \neq \mathcal{F}$.

12. Show that if μ is a finite measure, there cannot be uncountably many disjoint sets A with $\mu(A) > 0$.

1.3 Extension of Measures

In 1.3.A we discussed the concept of length of a subset of $\mathbb{R}$. The problem was to find a function μ giving the length of each interval, say $\mu([a,b]) = b - a$, on a fairly large class of sets. If $\mathcal{F}_0$ is the class of finite disjoint unions of right-closed intervals, there is no problem extending μ to $\mathcal{F}_0$: if $I_1, \ldots, I_n$ are disjoint right-closed intervals we set $\mu\left[\bigcup_{i=1}^n I_i\right] = \sum_{i=1}^n \mu(I_i)$. The resulting set function on $\mathcal{F}_0$ is finitely additive, but countable additivity is not clear at this point. Even if we can prove countable additivity on $\mathcal{F}_0$, we still have the problem of extending μ to the maximal σ-field over $\mathcal{F}_0$, namely the Borel sets.

We are going to consider a generalization of the above problem. Instead of working only with length, we shall examine sets of the form given by $\mu(A) = F(b) - F(a)$, where F is an increasing right-continuous function from $\mathbb{R}$ to $\mathbb{R}$. The extension technique to be developed is not restricted to sets of the form $(a,b]$, however, so we shall prove a general result concerning the extension of a measure from a field $\mathcal{F}_0$ to the minimal σ-field over $\mathcal{F}_0$.

It will be convenient to consider finite measures at first, and nothing is lost if we normalize and work with probabilities instead.

1.3.1 *Lemma.* Let $\mathcal{F}_0$ be a field of subsets of a space Ω, and let P be a probability measure on $\mathcal{F}_0$. Suppose that for each $A \in \mathcal{F}_0$, including $A = \Omega$, and

increasing limit A, and $\mu(A_n) \uparrow \mu(A)$ if $A_n \uparrow A$ (including $A_n \uparrow \Omega$) and increasing to A along A any sequence $A_n \uparrow A$ if $\mu(A) < \infty$ then

$$\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \mu(A_n).$$

Thus if A_n and A_n' both increase to the same limit A, then

$$\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \mu(A_n').$$

Proof. If n is fixed, $A_n \cap A_m \uparrow A_n$ as $m \to \infty$, so that

$$\mu(A_n \cap A_m) \to \mu(A_n)$$

by (1.2.7)(i). But $\mu(A_n \cap A_m) \le \mu(A_m)$ by (1.2.5)(ii), so that

$$\mu(A_n) = \lim_{m\to\infty} \mu(A_n \cap A_m) \le \lim_{m\to\infty} \mu(A_m).$$

Let $n \to \infty$ to obtain the proof.

We are now ready for the first extension of μ to a larger class of sets.

1.3.2 Lemma. Let μ be a probability measure on the field $\mathcal{A}_0$. Let $\mathcal{A}$ be the collection of all limits of increasing sequences of sets in $\mathcal{A}_0$, that is, $A \in \mathcal{A}$ if there exist $A_n \in \mathcal{A}_0$, $n = 1, 2, \ldots$, such that $A_n \uparrow A$. (Note that $\mathcal{A}$ can also be described as the collection of all countable unions of sets in $\mathcal{A}_0$; see 1.2.1.)

Define μ on $\mathcal{A}$ as follows: If $A_n \in \mathcal{A}_0$, $n = 1, 2, \ldots$, $A_n \uparrow A$, let $\mu(A) = \lim_{n\to\infty} \mu(A_n)$; μ is well-defined by 1.3.1, and $\mu = \mu$ on $\mathcal{A}_0$. Then:

(a) $\mathcal{A} \supset \mathcal{A}_0$ and $\mu(\emptyset) = 0$, $\Omega \in \mathcal{A}$ and $\mu(\Omega) = 1$; $0 \le \mu(A) \le 1$ for all $A \in \mathcal{A}$.

(b) If $G_1, G_2, \ldots \in \mathcal{A}$, then $G_1 \cup G_2 \cup \cdots \in \mathcal{A}$, and

$$\mu(G_1 \cup G_2) = \mu(G_1) + \mu(G_2) - \mu(G_1 \cap G_2).$$

(c) If $G_1, G_2 \in \mathcal{A}$ and $G_1 \subset G_2$, then $\mu(G_1) \le \mu(G_2)$.

(d) If $G_n \in \mathcal{A}$, $n = 1, 2, \ldots$, and $G_n \uparrow G$,

then $G \in \mathcal{A}$ and $\mu(G_n) \to \mu(G)$.

Proof. (a) This is clear since $\mu = P$ on $\mathcal{A}_0$ and P is a probability measure.

(b) If $A_n \in \mathcal{A}_0$, $A_n \uparrow G_1$, $A_n' \in \mathcal{A}_0$, $A_n' \uparrow G_2$, we have $\mu(A_n \cup A_m') = P(A_n \cup A_m') = P(A_n) + P(A_m') - P(A_n \cap A_m')$ by 1.2.5(ii). Let $n, m \to \infty$ to complete the argument.

(c) This follows from (b).

1.21 Since G is a countable union of sets in $\mathcal{A}$, $G = \bigcup_{n=1}^{\infty} G_n$, $G_n \in \mathcal{A}$. Show that each G_n can be written as $G_n = \bigcup_{k=1}^{\infty} A_{nk}$, $A_{nk} \in \mathcal{A}$, with $A_{nk} \uparrow G_n$ as $k \to \infty$. The situation may be represented schematically as follows:

$$\begin{array}{cccccc} A_{11} & A_{12} & A_{13} & \cdots & \uparrow & G_1 \\ A_{21} & A_{22} & A_{23} & \cdots & \uparrow & G_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \uparrow \\ A_{n1} & A_{n2} & A_{n3} & \cdots & \uparrow & G_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

Let $B_n = A_{11} \cup A_{22} \cup \cdots \cup A_{nn}$ (the B_n form an increasing sequence). The key step in the proof is the observation that

$$A_{nk} \subset B_n \subset G_n \qquad \text{for} \qquad k \le n \le m \tag{1}$$

and, therefore,

$$P(A_{nk}) \le P(B_n) \le \mu(G_n) \qquad \text{for} \qquad k \le n \le m. \tag{2}$$

Letting $k \to \infty$ in (2) we obtain $G_n = \bigcup_{k=1}^{\infty} A_{nk} \subset B_n \subset G_n$; then letting $n \to \infty$ we conclude that $B_n \uparrow G$, hence $P(B_n) \to \mu(G)$ by definition of μ. Now let $m \to \infty$ in (2) to obtain $\mu(G_n) \le \liminf_{k \to \infty} P(A_{nk}) \le \limsup_{k \to \infty} P(A_{nk}) \le \mu(G_n)$. Finally let $n \to \infty$ to conclude that $\lim_{k \to \infty} P(A_{nk}) = \lim_{n \to \infty} \mu(G_n) = \mu(G)$. ∎

We now extend μ to the class of all subsets of Ω. Henceforth, the extension will not be countably additive on all subsets, but only on a smaller σ-field. The construction depends on properties (1)–(3) of Lemma 1.2 and not on the fact that μ was derived from a probability measure on a field. We restate the result for easy reference.

1.22 Lemma. Let $\mathcal{A}$ be a class of subsets of a set Ω, μ a nonnegative real-valued set function on $\mathcal{A}$ such that $\mathcal{A}$ and μ satisfy the four conditions of Lemma 1.2. Define, for each $A \subset \Omega$,

$$\mu^*(A) = \inf\{\mu(G) : G \supset A,\ G \in \mathcal{A}_\sigma\}.$$

Then:

(a) $\mu^* = \mu$ on $\mathcal{A}$; $0 \le \mu^*(A) \le 1$ for all $A \subset \Omega$.

(b) If $A \subset B \subset \Omega$, then $\mu^*(A) \le \mu^*(B)$; in particular, $\mu^*(A) + \mu^*(A^c) \ge \mu^*(\Omega) = \mu(\Omega) = 1$.

(c) $\mu^*(A \cup B) \le \mu^*(A) + \mu^*(B)$, in particular, $\mu^*(A) + \mu^*(B) = \mu^*(A \cup B)$ if $AB = \emptyset$.

for $E \in \mathcal{E}$, then $\mu^*(E) = \mu(E)$.

(ii) If $A_n \uparrow A$, then $\mu^*(A_n) \uparrow \mu^*(A)$.

PROOF. (i) This is clear from the definition of μ^* and from 1.3.2(v).

(ii) If $\alpha < \mu^*(A)$, choose $G_1, G_2 \in \mathcal{S}_\sigma$, $G_1 \supset A_1$, $G_2 \supset A_2$, such that $\mu(G_1) < \mu^*(A_1) + \alpha/2$, $\mu(G_2) \le \mu^*(A_2) + \alpha/2$. By 1.3.2(iv),

$$\mu^*(A) = \mu^*(B) = \mu(G_1 \cup G_2) = \mu(G_1) + \mu(G_2) - \mu(G_1 \cap G_2) < \mu(G_1 \cap G_2)$$
$$< \mu^*(A_1) + \mu^*(A_2) + \alpha.$$

Since α is arbitrary, the result follows.

(iii) This follows from the definition of μ^*.

(iv) By (iii), $\mu^*(\bigcup_{n=1}^\infty A_n) \le \sum_{n=1}^\infty \mu^*(A_n)$. If $\alpha > 0$, for each n we may choose $G_n \in \mathcal{S}_\sigma$, $G_n \supset A_n$, such that

$$\mu(G_n) \le \mu^*(A_n) + \alpha 2^{-n}.$$

Now $A = \bigcup_{n=1}^\infty A_n \subset \bigcup_{n=1}^\infty G_n \in \mathcal{S}_\sigma$, hence

$$\mu^*(A) \le \mu\left(\bigcup_{n=1}^\infty G_n\right) \qquad \text{by (ii)}$$
$$= \mu\left(\bigcup_{n=1}^\infty G_n\right) \qquad \text{by (ii)}$$
$$= \lim_{N\to\infty} \mu\left(\bigcup_{n=1}^N G_n\right) \qquad \text{by 1.3.2(d)},$$

The proof will be accomplished if we prove that

$$\mu\left(\bigcup_{n=1}^N G_n\right) \le \sum_{n=1}^N \mu^*(A_n) + \alpha\left(1 - 2^{-N}\right) \le \sum_{n=1}^N \mu^*(A_n) + \alpha \qquad N = 1, 2, \ldots.$$

This is true for $N = 1$ by choice of G_1. If it holds for a given N, we apply 1.3.2(b) to the sets $\bigcup_{n=1}^N G_n$ and G_{N+1} to obtain

$$\mu\left(\bigcup_{n=1}^{N+1} G_n\right) = \mu\left(\bigcup_{n=1}^N G_n\right) + \mu(G_{N+1}) - \mu\left(\bigcup_{n=1}^N G_n \cap G_{N+1}\right).$$

Now $\{A_j\}_{j\in\mathbb{N}}$ is disjoint, $\bigcup_{j=1}^{n+1} A_j = G_{n+1}$, $\bigcup_{j=1}^{n} A_j = G_n$, so that the countable additivity yields

$$\mu^*\left(\bigcup_{j=1}^{n+1} A_j\right) = \mu^*(G_{n+1}) + \mu^*(A_{n+1}) - \mu^*(G_n) + \mu^*(A_{n+1}) = \mu^*(G_n)$$

$$\geq \mu^*(G_{n-1}) + c\sum_{j=1}^{n} 2^{-j} \geq \cdots \geq c\sum_{j=1}^{n} 2^{-j}.$$

Our aim in this section is to prove that a finite measure on a field $\mathcal{F}_0$ has a unique extension to the minimal σ-field over $\mathcal{F}_0$. That such an extension exists is not necessarily unique (in proving this more general result (see Problem 8) the following concept plays a key role:

1.3.4 **Definition.** An outer measure on Ω is a nonnegative, extended real-valued set function λ on the class of all subsets of Ω satisfying

(a) $\lambda(\emptyset) = 0$;
(b) $A \subset B$ implies $\lambda(A) \leq \lambda(B)$ (monotonicity); and
(c) $\lambda\left(\bigcup_{j=1}^{\infty} A_j\right) \leq \sum_{j=1}^{\infty} \lambda(A_j)$ (countable subadditivity).

The next theorem gives a unique outer measure on Ω. Parts (a)–(c) of it follow from 1.3.2(a), 1.3.2(b), and 1.3.3(a), and 1.3.4(c) is proved as follows:

$$\mu^*\left(\bigcup_{j=1}^{\infty} A_j\right) = \lim_{n\to\infty} \mu^*\left(\bigcup_{j=1}^{n} A_j\right) \qquad \text{by 1.3.3(b)};$$

$$\leq \lim_{n\to\infty} \sum_{j=1}^{n} \mu^*(A_j) \qquad \text{by 1.3.2(b)},$$

as desired.

We now identify a σ-field on which μ^* is countably additive:

1.3.5 **Theorem.** Under the hypotheses of 1.3.2, with μ^* defined as in 1.3.3, let
$\mathcal{M} = \{M \subset \Omega : \mu^*(M) + \mu^*(M^c) = 1\}$
[$M \in \mathcal{M}$ if $\mu^*(A) \geq \mu^*(A \cap M) + \mu^*(A \cap M^c)$ for every $A \subset \Omega$].
Then $\mathcal{M}$ is a σ-field and μ^* is a probability measure on $\mathcal{M}$.

Proof. First note that $\mathcal{F}_0 \subset \mathcal{M}$. For if $A_0 \in \mathcal{F}_0$ and $A \subset \Omega$, $\varepsilon > 0$, there is $\{B_j\} \subset \mathcal{F}_0$ with $A \subset \bigcup_j B_j$ and $\sum_j \mu^*(B_j) < \mu^*(A) + \varepsilon$. Then $\mu^*(A) + \varepsilon > \sum_j \mu^*(B_j \cap A_0) + \sum_j \mu^*(B_j \cap A_0^c) \geq \mu^*(A \cap A_0) + \mu^*(A \cap A_0^c)$, by 1.3.2(b), $\mu^*(A_0) + \mu^*(A_0^c) = 1$.

Clearly $\mathscr{S}'$ is closed under complementation, and $\Omega \in \mathscr{S}'$ by 1.3.2(iv) and 1.3.2(ii). If $B_1, B_2 \in \mathscr{S}'$, then by 1.3.3(iii),

$$\mu^*(A) \geq \mu^*(A \cap B_1) + \mu^*(A \setminus B_1) \geq \mu^*(A \cap B_1 \cap B_2) + \mu^*(A \setminus B_1 \cap B_2) + \mu^*(A \setminus B_1 \setminus B_2) \tag{1}$$

and since

$$(B_1 \cup B_2) \cap B_1 = B_1 \cup (B_2 \cap B_1), \qquad (B_1 \cup B_2) \setminus B_1 = B_2 \setminus B_1,$$

we have

$$\mu^*(A \cap (B_1 \cup B_2)) + \mu^*(A \setminus (B_1 \cup B_2)) \leq \mu^*(A \cap B_1) + \mu^*(A \setminus B_1). \tag{2}$$

If $\mu^*(A) = \infty$ for some $A \subset \Omega$, then $\mu^*(A \cap B) < \infty$ for at least one B by 1.3.2(iv), and the sum of the right sides of (1) exceeds the sum of the left sides as well. If $a = \mu^*(A \cap B_1 \cap B_2)$, $b = \mu^*(A \setminus B_1 \setminus B_2)$, $c = \mu^*(A \cap B_1 \setminus B_2)$, $d = \mu^*(A \setminus B_1 \cap B_2)$, then $a + b + c + d < \infty$, hence $a < \infty$, $b < \infty$, $c < \infty$, $d < \infty$. If $a = \infty$, then $b = c = d = 0$ since $\mu^*(A) < \infty$. Consequently $B_1 \cap B_2 \in \mathscr{S}'$ and $B_1 \cup B_2 \in \mathscr{S}'$. We have therefore shown that $\mathscr{S}'$ is a field. Now equality must hold in (1), for if not, the sum of the left sides of (1) and (2) would be less than the sum of the right sides, a contradiction. Thus μ^* is finitely additive on $\mathscr{S}'$.

To show that $\mathscr{S}'$ is a σ-field, let $B_n \in \mathscr{S}'$, $n = 1, 2, \ldots$, $A_n = \bigcup_{i=1}^{n} B_i$, $p(B_j) = \mu^*(B_j \cap A_n)$, $B = \bigcup_{i=1}^{\infty} B_i$. By 1.3.3(iii), hence by (1), $p(B_j) + \mu^*(B_j \setminus A_n) \geq \mu^*(B_j)$. Since $\mu^*(B_j \setminus A_n) \leq \mu^*(B_j)$ for all n by 1.3.3(ii), and $B_j \subset A_n$, we have $p(B_j) + \mu^*(B_j \setminus A_n) \leq \mu^*(B_j)$. Therefore p is additive, $B \in \mathscr{S}'$, making $\mathscr{S}'$ a σ-field.

Since $\mu^*(B_n \cap \cdot) = \mu^*(B_n \cap \cdot)$, μ^* is countably additive by 1.3.5(iv). $\square$

We now have our first extension theorem.

1.3.6 **Theorem.** *A finite measure on a field $\mathscr{S}$ can be extended to a measure on $\sigma(\mathscr{S})$.*

Proof. Nothing is lost by considering a probability measure. (Replace μ by $\mu/\mu(\Omega)$ if necessary.) The extension is then unique from 1.3.4 if $\mathscr{S}$ is countably additive. If $\mathscr{S} \subset \mathscr{S}' \subset \mathscr{M}$, where $\sigma(\mathscr{S}) \subset \mathscr{M}$, then μ^* restricted to $\sigma(\mathscr{S})$ is the desired extension. $\square$

In fact there is very little difference between $\sigma(\mathscr{S})$ and $\mathscr{S}$: if $A \in \sigma(\mathscr{S})$, then A can be expressed as $A = N_1 \setminus N_2$, where $N_1 \in \sigma(\mathscr{S})$ and N_2 is a subset of a set $M \in \sigma(\mathscr{S})$ with $\mu^*(M) = 0$. To establish this we introduce the idea of simple versus inessential extension:

1.3.7 **Definitions.** *A measure μ on a σ-field $\mathscr{S}'$ is said to be complete iff whenever $A \in \mathscr{S}'$ and $\mu(A) = 0$ we have $B \in \mathscr{S}'$ for all $B \subset A$.*

In 1.3.3, μ^* and ν^* is countable, or if $S = \{A \subset \mathcal{M} : \mu^*(A) = 0\}$, then $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \setminus A) \geq \mu^*(B \cap A) + \mu^*(B \setminus A) = \mu^*(B)$, thus $B \in \mathcal{M}$.

The completion of a measure space $(\Omega, \mathcal{F}, \mu)$ is defined as follows. Let $\mathcal{N}_0$ be the class of sets $A \subset N$ where N ranges over $\mathcal{F}$ and $\mathcal{N}$ over all subsets of sets of measure 0 in $\mathcal{F}$.

Now $\mathcal{F}_0$ is a σ-field including $\mathcal{F}$. For if $A \in \mathcal{F}$, $N \subset M \in \mathcal{F}$, $\mu(M) = 0$, then $A \cap N = (\Omega \setminus A) \cap M \in \mathcal{F}$, $\mu(A \cap N) \leq \mu(M) = 0$, and $A \setminus N = A \cap (\Omega \setminus N) = A \setminus (A \cap N) \in \mathcal{F}$, since $A \cap N \in \mathcal{F}$. Thus $A \setminus N \in \mathcal{F}_0$.

We extend μ to $\mathcal{F}_0$ by setting $\mu(A \cup N) = \mu(A)$. This is a valid definition, for if $A_1 \cup N_1 = A_2 \cup N_2$, $N_i \subset M_i \in \mathcal{F}$, $\mu(M_i) = 0$, then

$$\mu(A_1) \leq \mu(A_1 \cup N_1) \leq \mu(A_1) + \mu(M_2) = \mu(A_1 \cup M_2) \leq \mu(A_2) + \mu(M_2) = \mu(A_2),$$

since $A_1 \cup M_2 \supset A_2 \cup N_2$. Thus $\mu(A_1) \leq \mu(A_2)$, and by symmetry, $\mu(A_1) = \mu(A_2)$. The measure space $(\Omega, \mathcal{F}_0, \mu)$ is called the completion of $(\Omega, \mathcal{F}, \mu)$, and $\mathcal{F}_0$ the completion of $\mathcal{F}$ relative to μ.

Note that the completion is in fact complete, for if $A \subset A_1 \cup N \in \mathcal{F}_0$, where $A_1 \in \mathcal{F}$, $N \subset M \in \mathcal{F}$, $\mu(M) = 0$, then $A \subset A_1 \cup M \in \mathcal{F}$, $\mu(A_1 \cup M) = 0$, thus $A \in \mathcal{F}_0$.

2.3.2 Theorem. In 1.3.6, $(\Omega, \mathcal{M}, \mu^*)$ is the completion of $(\Omega, \mathcal{F}, \mu^*)$.

Proof. We must show that $\mathcal{M} = \mathcal{F}_0$, where $\mathcal{F} = \sigma(\mathcal{E})$. If $A \subset B \in \mathcal{F}$, $\mu^*(B) = 0$, and $\mu^*(A) > 0$, we can find sets $G_n \in \sigma(\mathcal{E})$, $A \subset G_n \subset B$, and $\mu^*(G_n) < \mu^*(A) + 1/n$. Let $G = \bigcap_1^\infty G_n$, $G' = \bigcap_1^\infty G_n'$. Then $A = G \setminus (G \setminus A)$, $G \in \sigma(\mathcal{E})$, $G \setminus A \subset B \in \sigma(\mathcal{E})$, $\mu^*(G \setminus A) \leq \mu^*(G_n \setminus A) \leq \mu^*(G_n) - \mu^*(A) < 1/n$, so that $\mu^*(G \setminus A) = 0$. Thus $A \in \mathcal{F}_0$.

Conversely, if $A \in \mathcal{F}_0$, then $A = A_1 \cup N$, $A_1 \in \mathcal{F}$, $N \subset M \in \mathcal{F}$, $\mu^*(M) = 0$. Since μ^* is σ-finite on $\mathcal{E}$, $A_1 \in \mathcal{M}$, and since $(\Omega, \mathcal{F}, \mu^*)$ is complete we have $N \in \mathcal{M}$. Thus $A \in \mathcal{M}$.

To prove the uniqueness of the extension from $\mathcal{F}_0$ to $\mathcal{F}$, we need the following basic result:

2.3.3 Monotone Class Theorem. Let $\mathcal{F}_0$ be a field of subsets of Ω, and $\mathcal{C}$ a class of subsets of Ω that is monotone (if $A_n \subset A_{n+1}$ and $A_n \in \mathcal{C}$, then $\bigcup_1^\infty A_n \in \mathcal{C}$; if $A_n \supset A_{n+1}$ and $A_n \in \mathcal{C}$, then $\bigcap_1^\infty A_n \in \mathcal{C}$). If $\mathcal{C} \supset \mathcal{F}_0$, then $\mathcal{C} \supset \sigma(\mathcal{F}_0)$, the minimal σ-field over $\mathcal{F}_0$.

Proof. The technique of the proof is called "back stepping". Let $\mathcal{M} = m(\mathcal{F}_0)$ and let $m(\mathcal{F}_0)$ be the smallest monotone class containing all sets of

...We show that $\mathcal{M} = \mathcal{M}_1$ in order to...

Fix $A \in \mathcal{S}$ and let $\mathcal{N}_A = \{B \in \mathcal{M} : A \cap B \in \mathcal{M}\}$; then $\mathcal{N}_A$ is a monotone class. For $A \in \mathcal{S}_0$ or if $A \in \mathcal{M}_1$, then $\mathcal{S} \subset \mathcal{N}_A$ since $\mathcal{S}_0$ is a field. Hence $\mathcal{M} \subset \mathcal{N}_A$ by minimality of $\mathcal{M}$, consequently $\mathcal{M} = \mathcal{N}_A$. But this shows that for any $B \in \mathcal{M}$ we have $A \cap B, A \cap B^c, A \cup B \in \mathcal{M}$ whenever $A \in \mathcal{S}_0$ or $A \in \mathcal{M}_1$, so that $\mathcal{M} = \mathcal{M}_1$. Again by minimality of $\mathcal{M}$, $\mathcal{M}_1 = \mathcal{M}$.

Now $\mathcal{M}$ is a field so that $A, B \in \mathcal{M} \Rightarrow A \cup B \in \mathcal{M}$. Then $A \cap B, A \cap B^c, A^c \cap B$ and $A \cup B$ are pairwise disjoint sets in $\mathcal{M}$ and a monotone class argument also shows that $\mathcal{M}$ is a σ-field because $\mathcal{M}$ is a monotone class including $\mathcal{S}_0$.

We now prove the fundamental extension theorem.

1.3.10 Caratheodory Extension Theorem. Let μ_0 be a measure on the field $\mathcal{S}_0$ of subsets of Ω and assume that μ_0 is σ-finite on $\mathcal{S}_0$, so that Ω can be decomposed as $\bigcup_1^\infty A_n$, where $A_n \in \mathcal{S}_0$ and $\mu_0(A_n) < \infty$ for all n. Then μ_0 has a unique extension to a measure on the minimal σ-field $\mathcal{S}$ over $\mathcal{S}_0$.

Proof. Since $\mathcal{S}_0$ is a field, the A_n may be taken as disjoint (replace A_n by $A_n \cap A_{n-1}^c \cap \cdots \cap A_1^c$). Let us assume (1) of 1.1. Let $\mu_0(A) = \mu(A \cap \cdots \cap A_n)$, $A \in \mathcal{S}_0$; then μ_n is a finite measure on $\mathcal{S}_0$, hence by 1.3.6 it has an extension μ_n^* on $\mathcal{S}$. Also $\mu_n \uparrow \mu_0$ so that the function $\mu^* = \sum_{n=1}^\infty \mu_n^*$ is a measure on $\mathcal{S}$ since the order of summation is only countably many finite quantities can be interchanged.

Now suppose that ν is a measure on $\mathcal{S}$ and $\nu = \mu_0$ on $\mathcal{S}_0$. Then $\nu(A \cap A_n) = \mu(A \cap A_n)$ $= \mu_n(A)$, $A \in \mathcal{S}$. Since μ_n is a finite measure on $\mathcal{S}$ and $\mu_n = \nu_n = \mu_0$ on $\mathcal{S}_0$, and $\mathcal{S}_0$ is a field, $\mu_n = \mu_n^*$ on $\mathcal{S}$. For $A \in \mathcal{S}$, $\mu(A) = \lim_n \mu(A \cap A_n) = \mu^*(A)$ by monotone convergence (1.2.1). Thus ν coincides with μ^* on $\mathcal{S}$. Hence $\mu^* = \nu$ by 1.3.9. But then $\mu = \sum_n \mu_n = \sum_n \mu_n^* = \mu^*$, proving uniqueness. □

The intuitive idea of constructing a maximal σ-field by countable unions, complements and countable unions and intersections in all possible ways suggests that if $\mathcal{M}_1$ is a field on Ω and $\mathcal{M}$ is the σ-field over $\mathcal{M}_1$, then sets in $\mathcal{M}$ can be approximated in some sense by sets in $\mathcal{M}_1$. The following result crystallizes this notion.

1.3.11 Approximation Theorem. Let $(\Omega, \mathcal{M}, \mu)$ be a measure space, and let $\mathcal{M}_1$ be a field of subsets of Ω such that $\mathcal{M} = \mathcal{M}_1$. Assume that μ is σ-finite on $\mathcal{M}_1$ and let $\varepsilon > 0$ be given. If $A \in \mathcal{M}$ and $\mu(A) < \infty$, there is a set $B \in \mathcal{M}_1$ such that $\mu(A \triangle B) < \varepsilon$.

Proof. Let $\mathcal{V}$ be the class of all countable unions of sets of $\mathcal{F}_0$. The conclusion of 1.3.11 holds for any $A \in \mathcal{V}$, by 1.2.7(a) of 1.3.3, if μ is finite and $A \in \mathcal{F}_0$, it can be approximated externally closely (in the sense of 1.3.11) by a set in $\mathcal{F}_0$, and the same 1.3.11 is proved for finite μ. In general, let Ω be the disjoint union of sets $A_\xi \in \mathcal{V}$ with $\mu(A_\xi) < \infty$, and let $\nu(E) = \mu(E \cap A_\xi)$, $E \subseteq X$.

Then μ_ξ is a finite measure on $\mathcal{P}X$, hence is in $\mathcal{C}\mathcal{P}$. Take a set $B_\xi \in \mathcal{F}_0$ such that $\mu(B_\xi \triangle E_\xi) < \varepsilon 2^{-\xi}$. Since

$$\mu_\xi(E \triangle B_\xi) = \mu(E \triangle B_\xi) \cap A_\xi)$$
$$= \mu[(E \cap A_\xi) \cup (B_\xi \cap A_\xi)] - \mu[(E \cap A_\xi) \cap (B_\xi \cap A_\xi)],$$

and $B_\xi \cap A_\xi \in \mathcal{F}_0$, we may assume that $B_\xi \subseteq A_\xi$. (The observation that $B_\xi \cap A_\xi \in \mathcal{F}_0$ is the point where we use the characterization that $\mathcal{F}_0$ is the saturated algebra generated by $\mathcal{F}$.) Then if $G = \bigcup_\xi B_\xi$, then $G \in \mathcal{F}_0$ and

$$\mu_\xi(G \cap C) = \mu_\xi(B_\xi \cap C) \text{ for } A_\xi, \quad \mu(G \cap C) = \mu(B_\xi \cap B_\eta) \text{ for } A_\eta, \quad \mu_\xi = \mu_\xi \text{ for } B_\eta,$$

hence

$$\mu_\xi(G \cap C) = \sum_{\eta=1}^{\xi} \mu_\eta(B_\eta \cap C) + \mu_\xi(B_\xi \cap C) \text{ for } \bigcup_{\eta=1}^{\xi} A_\eta, \quad C = A_\xi \text{ for } \xi = \eta = \infty.$$

and $\mu(G \cap C) = \sum_{\xi=1}^{\infty} \mu(B_\xi \cap C)$ for $A \in \mathcal{F}$ and $\mu(G) < \infty$. It follows from 1.2.7 that $\mu(A \triangle G) < \varepsilon$ for $A \in \mathcal{F}$ and $\mu(A) < \infty$. If μ is σ-finite, then $G = \bigcup_{n=1}^{\infty} G_n$ where $G_n \in \mathcal{F}_0$.

2.3.22 Example. Let Ω be the rationals, $\mathcal{F}_0$ the field of finite disjoint unions of right-semiclosed intervals in $\mathbb{R}$ ($[a, b) = \{x : a \le x < b\}$, a, b rational, including $a = -\infty$) and Ω itself is right semiclosed, see 1.2.2. Let $\mathcal{F} = \sigma(\mathcal{F}_0)$. Then:

(a) $\mathcal{F}_0$ consists of all subsets of Ω.

(b) If $\mu(\Omega)$ is the number of points in Ω (μ is counting measure), then μ is σ-finite on $\mathcal{F}$ but not on $\mathcal{F}_0$.

(c) There are sets $A \in \mathcal{F}$ of finite measure that cannot be approximated by sets in $\mathcal{F}_0$, that is, there is no sequence $A_n \in \mathcal{F}_0$ with $\mu(A \triangle A_n) \to 0$.

(d) If $\nu = \Delta_0 \mu$, then $\nu = \mu$ on $\mathcal{F}_0$ but not on $\mathcal{F}$.

Thus both the approximation theorem and the Carathéodory extension theorem fail in this case.

choice. (b) We have $|x| = \lim_n |x_n| \le \alpha$. Hence $x \in [-\alpha, \alpha]$, and therefore all singletons are in $\mathcal{F}$. But then all sets are in $\mathcal{F}$ since Ω is countable.

(c) Since Ω is countable by (a), each nonempty set in $\mathcal{F}$ has infinite measure, so it is not σ-finite on $\mathcal{F}$.

(d) If A is any finite nonempty subset of Ω, then $\mu(A) > 0$. However all subsets $B \subset A^c$, because any nonempty set in $\mathcal{F}$ must contain infinitely many points of Ω.

(e) Since $\lambda(\{x\}) = 0$ and $\mu(\{x\}) = 1$, $\lambda \ll \mu$ on $\mathcal{F}$. But $\lambda(A) = \inf\{\sum_{n=1}^{\infty} \lambda(\{x_n\}) : A \subset \bigcup_{n=1}^{\infty} \{x_n\}\} = 0$ (except for $A = \Omega$), $\mu(A) = 1$.

Problems

1. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $\mathcal{F}_0$ be the completion of $\mathcal{F}$ relative to μ. If $A \subset \Omega$, define

$$\mu_*(A) = \sup\{\mu(B) : B \subset A, B \in \mathcal{F}\} \quad \mu^*(A) = \inf\{\mu(B) : A \subset B, B \in \mathcal{F}\}.$$

   If $A \in \mathcal{F}_0$, show that $\mu_*(A) = \mu^*(A) = \mu(A)$. Conversely, if $\mu_*(A) = \mu^*(A) < \infty$, show that $A \in \mathcal{F}_0$.

2. Show that the monotone class theorem (1.3.9) fails if $\mathcal{F}_0$ is not assumed to be a field.

3. This problem deals with the extension of an arbitrary (not necessarily finite) measure on a field.

   (a) Let λ be an arbitrary measure on the sets of some field $\mathcal{F}_0$. We say that the set E is measurable if

$$\lambda(A) = \lambda(A \cap E) + \lambda(A \cap E^c) \quad \text{for all} \quad A \in \mathcal{F}_0.$$

   (The inequality may be replaced by "$\le$" by subadditivity of λ.) If $\mathcal{F}$ is the class of all measurable sets, show that $\mathcal{F}$ is a field, and that if $E_1, E_2, \ldots$ are disjoint sets in $\mathcal{F}$ with union E, and $A \subset \Omega$, we have

$$\lambda(A \cap E) = \sum_{n=1}^{\infty} \lambda(A \cap E_n) \tag{1}$$

   In particular, λ is a measure on $\mathcal{F}$. (Use the definition of measurability to show that $\mathcal{F}$ is a field and that (1) holds for finite sequences. If $E_1, E_2, \ldots$ are disjoint sets in $\mathcal{F}$ and $F_n = \bigcup_{i=1}^{n} E_i \in \mathcal{F}$, show that

$$\lambda(A \cap F_n) = \lambda(A \cap F_n \cap E_n) + \lambda(A \cap F_n \cap E_n^c) = \lambda(A \cap E_n) + \lambda(A \cap F_{n-1})$$

   and then let $n \to \infty$.

(b) Let μ be a measure on a field $\mathcal{A}_0$ of subsets of Ω. If $A \subset \Omega$, define

$$\mu^*(A) = \inf\left\{\sum_n \mu(A_n) : A \subset \bigcup_n A_n,\ A_n \in \mathcal{A}_0\right\}.$$

Show that μ^* is an outer measure on 2^Ω and that $\mu^* = \mu$ on $\mathcal{A}_0$.

(c) In part (b), if $\mathcal{A}_0$ is the class of finite unions of half-open intervals $(a, b]$, show that $\mathcal{B}_1 \subset \mathcal{M}^*$. Conclude by (b) and (a) that μ may be extended to the Lebesgue σ-field over $\mathbb{R}$.

(d) In (b), if μ is finite on $\mathcal{A}_0$, show that $\mathcal{M}^*$ is the completion of $\mathcal{B}_1$ with respect to μ^*.

1.4 Lebesgue–Stieltjes Measures and Distribution Functions

We are now in a position to construct a large class of measures on the Borel sets of $\mathbb{R}$. If F is an increasing, right-continuous real function from $\mathbb{R}$ to $\mathbb{R}$, we let $\mu((a, b]) = F(b) - F(a)$. We then extend μ to a finitely additive set function on the field $\mathcal{A}_0$ of finite disjoint unions of right-semiclosed intervals. If we can show that μ is countably additive on $\mathcal{A}_0$, the Carathéodory extension theorem extends μ to $\mathcal{B}_1$.

1.4.1 Definition. A *Lebesgue–Stieltjes measure* on $\mathbb{R}$ is a measure μ on $\mathcal{B}_1$ such that $\mu(I) < \infty$ for each bounded interval I. A *distribution function* on $\mathbb{R}$ is a map $F : \mathbb{R} \to \mathbb{R}$ that is nondecreasing ($s < t$ implies $F(s) \le F(t)$) and right-continuous ($\lim_{t \to s^+} F(t) = F(s)$). We are going to show that for each Lebesgue–Stieltjes measure μ, $F(t) = \mu((0, t])$ $(t \ge 0)$ sets up a one-to-one correspondence between Lebesgue–Stieltjes measures and distribution functions that vanish at 0, where two distribution functions differ by a constant are identified.

1.4.2 Proposition. Let μ be a Lebesgue–Stieltjes measure on $\mathbb{R}$. Then $F : \mathbb{R} \to \mathbb{R}$ is defined up to an additive constant by $F(t) - F(s) = \mu((s, t])$. [For example, let $F(t)$ arbitrarily and set $F(s) = F(0) - \mu((s, 0])$, $s < 0$; $F(t) = F(0) + \mu((0, t])$, $t > 0$.] Then F is a distribution function.

Proof. If $s < t$, then $F(t) - F(s) = \mu((s, t]) \ge 0$. If $\{x_n\}$ is a sequence of points such that $x_n \to x$ and $x_n > x$, then $F(x_n) - F(x) = \mu((x, x_n])$. If $x_n \downarrow x$, then $(x, x_n] \downarrow \emptyset$, so $F(x_n) \to F(x)$. $\square$

Now let F be a nondecreasing right-continuous function on $\mathbb{R}$. It will be convenient to work in the compact space $\bar{\mathbb{R}}$, so we extend F to a map of $\bar{\mathbb{R}}$ into $\bar{\mathbb{R}}$ by defining $F(\infty) = \lim_{x \to \infty} F(x)$, $F(-\infty) = \lim_{x \to -\infty} F(x)$, the limits existing by monotonicity. Define $\mu((a, b]) = F(b) - F(a)$ if $-\infty \le a < b \le \infty$, and

for $g(-\infty,0] = g^+(0^+) - g^+(-\infty) = g^+(-\infty,0]$; and μ is defined on all right semi-closed intervals of $\mathbb{R}$ (as $\mu(a,b] = \mu(-\infty,b] - \mu(-\infty,a]$, as in (1.2.2)).

If $I_1, I_2, \ldots, I_n$ are disjoint right semi-closed intervals of $\mathbb{R}$, we define $\mu(\bigcup_{k=1}^n I_k) = \sum_{k=1}^n \mu(I_k)$. This gives an extension of the field $\mathcal{A}_0(\mathbb{R})$ of right semi-closed intervals and μ becomes additive on $\mathcal{A}_0(\mathbb{R})$. To show that μ is in fact countably additive on $\mathcal{A}_0(\mathbb{R})$, we make use of (1.2.5)(i), as follows.

1.4.3 *Lemma.* The set function μ is countably additive on $\mathcal{A}_0(\mathbb{R})$.

Proof. First we show that $F(x_0) - F(-\infty)$ for some x_0, so that μ is finite. Let $A_1, A_2, \ldots$ be a sequence of sets in $\mathcal{A}_0(\mathbb{R})$ decreasing to $\emptyset$. If no A_n is one of the intervals of A_n then by right continuity at x_0, $\mu(A_n) = F(b_n) - F(a_n) < \varepsilon/2^n$, as in the proof of Theorem 1.2.5.

Thus we can find sets $B_n \in \mathcal{A}_0(\mathbb{R})$ whose closures $\overline{B}_n$ (in $\mathbb{R}$) are included in A_n, with $\mu(B_n)$ approximating $\mu(A_n)$. There is no loss in assuming that the finitely many endpoints of the intervals in B_n are rational; $\mu(A_n - B_n) < \varepsilon/2^n$. Now $\bigcap_{n=1}^\infty \overline{B}_n = \emptyset$, and it follows that $\bigcap_{n=1}^k \overline{B}_n = \emptyset$ for some finitely large k. To see this in detail, note that for each n the open sets $\mathbb{R} - \overline{B}_n$ form an open covering of the compact set $\overline{B}_1$; hence there is a finite subcovering, so that $\bigcup_{n=1}^k (\mathbb{R} - \overline{B}_n) = \mathbb{R}$ for some k. Therefore $\bigcap_{n=1}^k \overline{B}_n = \emptyset$. Now

$$\begin{aligned}
\mu(A_1) &= \mu\left(A_1 - \bigcap_{n=1}^k B_n\right) + \mu\left(\bigcap_{n=1}^k B_n\right) \\
&= \mu\left(\bigcup_{n=1}^k (A_n - B_n)\right) + \mu\left(\bigcap_{n=1}^k B_n\right) \qquad \text{since } A_1 \supset A_2 \supset \cdots \supset A_k \\
&\le \sum_{n=1}^k \mu(A_n - B_n) + \mu\left(\bigcap_{n=1}^k B_n\right) \qquad \text{by (1.2.5)(i)} \\
&< \varepsilon.
\end{aligned}$$

Thus $\mu(A_n) \to 0$.

Now if $F(x_0) - F(-\infty) = \alpha$, say, then $\mu(A_n) = F(b_n) - F(a_n) \to \alpha$; $\mu(A_n) = F(b_n) - F(-\infty) \to \alpha$. Let μ_0 be the set function corresponding to $F_0(x) = F(x) - \alpha$. Then μ_0 is finite and $\mu_0(A_n) \to 0$ for disjoint sets in $\mathcal{A}_0(\mathbb{R})$ since $\mu_0(A_n) = \mu(A_n) - \alpha$. Therefore μ_0 is countably additive on $\mathcal{A}_0(\mathbb{R})$, so that $\mu_0(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu_0(A_n)$ if $A_n \in \mathcal{A}_0(\mathbb{R})$ are disjoint and $\bigcup_{n=1}^\infty A_n \in \mathcal{A}_0(\mathbb{R})$. Hence $\mu(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$.

(Problem) Section $\sum_{i=1}^{n} a_i \chi_{A_i}$ of $\sum_{i=1}^{n} a_i \chi_{A_i}$ is an integral of $\sum_{i=1}^{n} a_i \chi_{A_i}$ and

$$\int \sum_{i=1}^{n} a_i \chi_{A_i} = \sum_{i=1}^{n} a_i \int \chi_{A_i}$$

$$= \sum_{i=1}^{n} a_i \mu(A_i)$$

since the a_i are all non-zero. Thus $\int \sum_{i=1}^{n} a_i \chi_{A_i} = \sum_{i=1}^{n} a_i \mu(A_i)$. We may write

$$\int \sum_{i=1}^{n} a_i \chi_{A_i} = \sum_{i=1}^{n} a_i \mu(A_i)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \mu(A_i \cap B_j)$$

$$= \int \sum_{j=1}^{m} \sum_{i=1}^{n} a_i \chi_{A_i \cap B_j}$$

We now consider the construction of Lebesgue–Stieltjes measures.

2.4.4 *Theorem.* Let F be a distribution function on $\mathbb{R}$. Then there exists a unique measure μ on $\mathcal{B}(\mathbb{R})$ such that $\mu((a,b]) = F(b) - F(a)$ for all $a, b \in \mathbb{R}$ with $a < b$. The measure μ is called the Lebesgue–Stieltjes measure associated with F.

Proof. Consider the collection $\mathcal{A}$ of all finite disjoint unions of right-semi-closed intervals of $\mathbb{R}$ (including (a, ∞) as right-semi-closed, see 1.2.2) and $\mathbb{R}$ itself. It is easy to see that $\mathcal{A}$ is an algebra of subsets of $\mathbb{R}$ (see Problem 1.2.3). Define μ on $\mathcal{A}$ by $\mu((a,b]) = F(b) - F(a)$, $\mu(\mathbb{R}) = F(\infty) - F(-\infty) = 1$, $\mu(\emptyset) = 0$. Note that there is another possible choice for μ on $\mathbb{R}$, by $F(\infty) - F(-\infty)$. Now define

$$\mu((a,b]) = F(b) - F(a) \quad \text{if } a, b \in \mathbb{R} \text{ or } b = \infty$$

$$\mu((a,\infty)) = F(\infty) - F(a) \quad \text{if } a \in \mathbb{R}$$

and $\mu(\emptyset) = 0$. It is easy to check that μ is a one-to-one correspondence between a subset of $\mathcal{A}$ containing all sets of the form $(a,b]$ and $\mathcal{B}(\mathbb{R})$. It follows that μ is countably additive on $\mathcal{B}(\mathbb{R})$. Furthermore, μ is a finite measure on $\mathcal{B}(\mathbb{R})$ since $\mu(\mathbb{R}) = 1$. Also note that μ is not a finite measure on $\mathcal{B}(\mathbb{R})$ since the sets $(-n, n]$ do not cover $\mathbb{R}$. By the Carathéodory extension theorem, μ has a unique extension to $\mathcal{B}(\mathbb{R})$. The extension is a Lebesgue–Stieltjes measure because $\mu((a,b]) = F(b) - F(a)$. F is unique for each μ on $\mathcal{B}(\mathbb{R})$ and μ is

1.4.3 Comments and Examples. If F is a distribution function and μ the corresponding Lebesgue–Stieltjes measure, we have seen that $\mu((a,b]) = F(b) - F(a)$, $a < b$. The measure of any interval, right-or-left-closed or not, may be expressed in terms of F, but it $F(x-)$ denotes $\lim_{y \uparrow x} F(y)$, then

(1) $\mu([a,b]) = F(b) - F(a-)$ (2) $\mu((a,b)) = F(b-) - F(a)$,

(3) $\mu([a,b)) = F(b-) - F(a-)$, (4) $\mu(\{a\}) = F(a) - F(a-)$.

Thus F is continuous at a if and only if $\mu(\{a\}) = 0$. The expressions on the right in (1) – (4) are sufficient to prove (2). Observe that

$$\mu((a,b)) = \lim_{n \to \infty} \mu\left((a, b - \frac{1}{n}]\right) = \lim_{n \to \infty} F\left(b - \frac{1}{n}\right) - F(a) = F(b-) - F(a).$$

Statement (3) follows because

$$\mu([a,b)) = \lim_{n \to \infty} \mu\left((a - \frac{1}{n}, b)\right) = \lim_{n \to \infty} F(b-) - F\left(a - \frac{1}{n}\right) = F(b-) - F(a-).$$

(4) is proved similarly. The proof of (2) works even if $a = b$, so that $\mu(\{a\}) = F(a) - F(a-)$.

(5) If F is continuous at x and $\mu(\{x\}) = 0$, the magnitude of discontinuity of f at x coincides with the measure of $\{x\}$.

The following from the definition of F (3) by a limiting approach $-\infty$ or ∞ is important:

(6) $\mu((-\infty, a]) = F(a) - F(-\infty)$, (7) $\mu((a, \infty)) = F(\infty) - F(a)$,

(8) $\mu((-\infty, a)) = F(a-) - F(-\infty)$, (9) $\mu(\mathbb{R}) = F(\infty) - F(-\infty)$

(10) $\mu((a,\infty)) = F(\infty) - F(a)$.

(The notation in (6), (8), and (10) was already used in the proof of 1.4.1.)

It is to be noted that if μ is connected, since F may always be adjusted by an additive constant, nothing is lost in this case if we set $F(-\infty) = 0$.

We may now generate a large number of measures on $\mathcal{B}(\mathbb{R})$. The reasoning: If $f : \mathbb{R} \to \mathbb{R}$, $f \geq 0$, and f is integrable (Riemann for now) on any finite interval, then if we fix $a \in \mathbb{R}$ arbitrarily, we define

$$F(x) = F(a) + \int_a^x f(t)\,dt, \qquad x \geq a$$

$$F(x) = F(a) - \int_x^a f(t)\,dt, \qquad x < a$$

that F is a (nondecreasing) distribution function and thus gives rise to a Lebesgue–Stieltjes measure, namely

$$\mu(a,b] = \int_a^b F(t)\,dt.$$

In particular, we may take $g(x) = 1$ for all x, and $F(x) = x$, then $\mu(a,b] = b - a$. The set function μ is called the *Lebesgue measure* on $\mathcal{B}(\mathbb{R})$. The completion of $\mathcal{B}(\mathbb{R})$ relative to Lebesgue measure is called the class of *Lebesgue measurable sets*, written $\mathcal{L}(\mathbb{R})$. Thus a Lebesgue measurable set is the union of a Borel set and a subset of a Borel set of Lebesgue measure 0. The extension of Lebesgue measure on $\mathcal{L}(\mathbb{R})$ is also called "Lebesgue measure."

Now let μ be a Lebesgue–Stieltjes measure that is concentrated on a countable set $S = \{x_1, x_2, \ldots\}$ that is $\mu(S - S) = 0$. In general if $(S, \mathcal{S}, \mu)$ is a measure space and $A \in \mathcal{S}$, we say that μ is concentrated on A if $\mu(S - A) = 0$. In the present case, such a measure is called discrete. For $x \in S$, the magnitude $\mu\{x\}$ is the jump of μ at x. Since μ is a measure on all subsets of $\mathbb{R}$, not merely on the Borel sets, $\mu(A) = \sum_{x \in A} \mu\{x\}$ for all $A \subset S$. Moreover, since μ is finite on each bounded interval I, μ will be a Lebesgue–Stieltjes measure on $(\mathbb{R}, \mathcal{B})$, if $\sum_n \mu\{x_n\}$ converges, μ will be a finite measure. The distribution function F corresponding to μ is discontinuous on $\mathbb{R} - S$, if $\mu\{x\} > 0$, F has a jump of magnitude $\mu\{x\}$ at x. If $x < y$ and no point of S lies between x and y, then F is constant on $[x, y)$, but if $x \le x' < y$, then $F(x') - F(x) = \mu(x, x'] > 0$.

Now if we take S to be the set of rational numbers, this shows that a function F can be a monotone function from $\mathbb{R}$ to $\mathbb{R}$ and be continuous at each irrational point and discontinuous at each rational point.

If F is an increasing, right-continuous, real-valued function defined on a closed, bounded interval $[a, b]$, there is a corresponding finite measure μ on the Borel subsets of $[a, b]$, uniquely, as is determined by the requirement that $\mu(c, d] = F(d) - F(c)$ for $a \le c < d \le b$. The easiest way to establish the correspondence is to extend F by definition: $F(x) = F(a)$, $x < a$; $F(x) = F(b)$, $x \ge b$. One then takes as the Lebesgue–Stieltjes measure on intervals in $[a, b]$ precisely $\mu(c, d] = F(d) - F(c)$ for $a \le c < d \le b$.

We are going to consider Lebesgue–Stieltjes measures and distribution functions on Euclidean n-space. First, some terminology is required.

Definition and Comments. If $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, $x \le y$ means $x_i \le y_i$ for all $i = 1, \ldots, n$. The interval $(x, y]$ is defined as $(x, y] = \{(z_1, \ldots, z_n) \in \mathbb{R}^n : x_i < z_i \le y_i$ for all $i = 1, \ldots, n\}$. (x, y), $[x, y]$, $[x, y)$ are defined similarly. For all $i = 1, \ldots, n$, $(-\infty, y] = \{x \in \mathbb{R}^n : x_i \le y_i$ for all $i = 1, \ldots, n\}$, and other types

of intervals are defined similarly. The smallest σ-field containing all intervals $(a, b]$, $a, b \in \mathbb{R}^n$, is called the class of Borel sets of $\mathbb{R}^n$, and denoted $\mathcal{B}(\mathbb{R}^n)$. The Borel sets form the minimal σ-field over many other classes of sets. For example, the open sets, the intervals $[a, b)$, $[a, b]$, etc. all generate exactly as do the closed intervals $(a, b]$ in $\mathbb{R}^n$. The class of Borel sets of $\mathbb{R}^n$, written $\mathcal{B}(\mathbb{R}^n)$, is denoted similarly.

A *Lebesgue–Stieltjes measure* on $\mathbb{R}^n$ is a measure μ on $\mathcal{B}(\mathbb{R}^n)$ such that $\mu(B) < \infty$ for each bounded interval B.

The notion of a distribution function on $\mathbb{R}^n$, $n > 1$, is more complicated than in the one dimensional case. To see why, assume for simplicity that $n = 2$, and let μ be a finite measure on $\mathcal{B}(\mathbb{R}^2)$. Define

$$F(x_1, x_2) = \mu\{(a_1, a_2) \in \mathbb{R}^2 : a_1 \le x_1, \ a_2 \le x_2\}, \quad (x_1, x_2) \in \mathbb{R}^2.$$

By analogy with the one-dimensional case, we expect that F is a distribution function corresponding to μ, see Theorem 3.6 of Ch. 1 of [1]. This will turn out to be correct, but the correspondence is no longer one-to-one. In terms of the function $\mu(a, b] = F(b) - F(a)$. To see this, we compute $\mu(a, b]$ in terms of F instead of the difference operator Δ_h as follows.

If $G : \mathbb{R}^n \to \mathbb{R}$, $\Delta_{h_k, k} G(x_1, \ldots, x_n)$ is defined as

$$G(x_1, \ldots, x_k + h_k, \ldots, x_n) - G(x_1, \ldots, x_k, \ldots, x_n).$$

1.4.7 *Lemma.* If $a = (a_1, a_2)$, $b = (b_1, b_2)$, $a_i < b_i$, $i = 1, 2$, then

(a) $\mu(a, b] = \Delta_{h_1, 1} \Delta_{h_2, 2} F(x_1, x_2)$ where

(b) $\Delta_{h_1, 1} \Delta_{h_2, 2} F(x_1, x_2)$

$$= F(a_1 + h_1, a_2 + h_2) - F(a_1, a_2 + h_2) - F(a_1 + h_1, a_2) + F(a_1, a_2)$$

$$= \mu(a_1, a_1 + h_1] \times (a_2, a_2 + h_2] + \mu(a_1, a_1 + h_1] \times (a_2, a_2] + \mu(a_1, a_1] \times (a_2, a_2 + h_2]$$

where $h_i > 0$, $i = 1, 2$.

Proof.

(a) $\Delta_{h_1, 1} F(x_1, x_2) = F(x_1 + h_1, x_2) - F(a_1, x_2)$

$$= \mu\{(x, y) : x \le x_1 + h_1, \ y \le x_2\} - \mu\{(x, y) : x \le a_1, \ y \le x_2\}$$

$$= \mu\{(x, y) : a_1 < x \le x_1 + h_1, \ y \le x_2\}$$

$$= \mu\{(x, y) : a_1 < x \le x_1 + h_1, \ a_2 < y \le x_2\} + \mu\{(x, y) : a_1 < x \le x_1 + h_1, \ y \le a_2\}$$

since $a_2 \le x_2$.

similarly,

$$\Delta_{x_2,y_2}\Delta_{x_1,y_1}F(x,y,z)=\text{[illegible]}$$

and

$$\Delta_{x_3,y_3}\Delta_{x_2,y_2}\Delta_{x_1,y_1}F(x,y,z)=\text{[illegible]}$$

$$x_1<x_2<x_3.$$

(c) $\Delta_{x_1,y_1}F(x,y,z)=\text{[illegible]}$

$$\Delta_{x_2,y_2}\Delta_{x_1,y_1}\text{[illegible]}=\text{[illegible]}$$

$$=\text{[illegible]}$$

Thus $\Delta_{x_3,y_3}\Delta_{x_2,y_2}\Delta_{x_1,y_1}F(x,y,z)$ is the desired expression. ∎

The extension of 14.7 to n dimensions is clear.

14.8 *Theorem.* Let α be a finite measure on $\mathscr{B}(R^k)$. Then define

$$f(x)=\alpha(-\infty,x_1,\dots,x_k)\quad\text{for}\quad x=(x_1,\dots,x_k)\in R^k.$$

If $a<b$, then

(a) $\alpha([a,b])=\Delta_{b_1,\dots,b_k}\cdots\Delta_{a_1,\dots,a_k}f(x_1,\dots,x_k)$, where

(b) $\Delta_{x_k,\dots,x_1}\cdots\Delta_{x_2,\dots,x_k}f(x_1,\dots,x_k)=f_1-f_2+\cdots\pm f_k$, where f_j is the sum of all $\binom{k}{j}$ terms of the form $f(y_1,\dots,y_k)$ such that $y_i=x_i$ for exactly j indices in $[1,2,\dots,k]$, and $y_i=b_i$ for the remaining $k-j$ indices.

Proof. Apply 14.7 to the rectangle $[a,b]$ of 14.7. ∎

We know that a distribution function on R determines a correspondence, Lebesgue–Stieltjes measure. This is true in k dimensions. We shall prove the following theorem for increasing

Let $f:R^k\to R$, and, for $a=(a_1,\dots,a_k)$, $b=(b_1,\dots,b_k)$ denote

$$\Delta_{a_1,\dots,a_k}^{b_1,\dots,b_k}=\Delta_{x_k,\dots,x_1}^{y_k,\dots,y_1}.$$

The function f is said to be increasing if $\Delta_{a_1,\dots,a_k}^{b_1,\dots,b_k}f\ge 0$ whenever $a_i\le b_i$. f is right-continuous if it is right-continuous in all variables separately. In other words, for any sequence $x^1\ge x^2\ge\cdots\ge x^n\ge\cdots$ we have $f(x^n)\to f(x)$.

An increasing right continuous $F : \mathbb{R}^n \to \mathbb{R}$ is said to be a distribution function on $\mathbb{R}^n$. (Note that if F arises from a Lebesgue-Stieltjes measure as in 1.4.1, F is a distribution function.)

1.4.2 Given a distribution function on $\mathbb{R}^n$, we set $\mu((a,b]) = \Delta_{a,b}F$ (the notation as in 1.3.1). Then $\mu(\emptyset) = \lim_{x_1 \to -\infty} F(x) = 0$. We are going to show that μ has a unique extension to a Lebesgue-Stieltjes measure on $\mathbb{R}^n$. The technique of the proof is the same in any dimension, but to avoid cumbersome notation and to keep on the essential ideas, we summarize operations we need when $n = 2$. We break the argument into several steps.

(1) Let $\mathcal{A}$ be the union of the finite disjoint collections $\{(a_1,b_1] \times (a_2,b_2]\}$ formed by the rectangles in the first coordinate plane of the following three types:

$$(a_1, a_2] \times (a_2, \infty), \qquad (a_1, b_1] \times (b_2, \infty), \qquad (b_1, \infty) \times (a_2, b_2].$$

and their translations, the second coordinate of each of the following three types:

$$(a_1, \infty) \times (a_2, b_2], \qquad (a_1, b_1] \times (-\infty, a_2], \qquad (b_1, \infty) \times (-\infty, a_2].$$

For example, a typical set in the union is

$$\{(a_1, b_1] \times (a_2, \infty)\} \setminus \{(c_1, d_1] \times (c_2, \infty)\},$$

where the differences are taken in such sets.

Recall (1) may be verified by the following Fig. 1.4.1

(2) Let $\mathcal{A}(\mathbb{R}^1) = \bigcup_{n=1}^{\infty} \mathcal{I}(\mathbb{R}^1)$, where $\mathcal{I}(\mathbb{R}^1)$ denotes the semialgebra $\{(a,b] : -\infty \le a \le b < \infty\}$ of $\mathbb{R}^1$.

This is verified by induction using 1.4.1.

Now a right semiclosed interval $(a,b]$ in $\mathbb{R}^n$ is, by convention, a set of the form $\{(x_1, \ldots, x_n) : a_i < x_i \le b_i, i = 1, \ldots, n\}$, $a, b \in \mathbb{R}^n$, with the proviso that $a_i < x_i \le b_i$ can be replaced by $a_i \le x_i \le b_i$ if $a_i = -\infty$. With this assumption, the semialgebra $\mathcal{I}(\mathbb{R}^1)$ of finite disjoint unions of right semiclosed intervals is a lattice. The corresponding convention in $\mathbb{R}^n$ is that $a_i < x_i \le b_i$ can be replaced by $a_i \le x_i \le b_i$ if $a_i = -\infty$. Both conventions are dictated by consistency arguments in the proofs of the one dimensional case: see 1.3.2.)

(3) If $\mathcal{A}$ and $\mathcal{B}$ belong to $\mathcal{A}^*$ but not to $\mathcal{A}^*$, we define $\mathcal{A} \cap \mathcal{B}$ as the limit of $\mathcal{A}' \cap \mathcal{B}$ where $\mathcal{A}' \in \mathcal{I}(\mathbb{R}^n)$ decreases to $\mathcal{A}$ and $\mathcal{A}' \cap \mathcal{B}$ increases to $\mathcal{A} \cap \mathcal{B}$. (The definition is sensible because of the monotonicity property in 1.3.1.) Similarly if $\mathcal{A} \in \mathcal{I}(\mathbb{R}^n)$, $\mathcal{B} \in \mathcal{I}(\mathbb{R}^m)$, we take $F(\mathcal{A} \times \mathcal{B}) = \lim_{\mathcal{A}' \to \mathcal{A}} F(\mathcal{A}' \times \mathcal{B})$, if $\mathcal{A} \in \mathcal{I}(\mathbb{R}^n) \setminus \mathcal{I}(\mathbb{R}^n)$, $\mathcal{B} \in \mathcal{I}(\mathbb{R}^m)$. And $F(\mathcal{A} \times \mathcal{B}) = \lim_{\mathcal{B}' \to \mathcal{B}} F(\mathcal{A} \times \mathcal{B}')$.

The next step is to enlarge semiclosed intervals of $\mathbb{R}^1$ to extend an finitely additive set function on $\mathcal{I}(\mathbb{R}^1)$ as in the discussion after 1.4.2. There is a

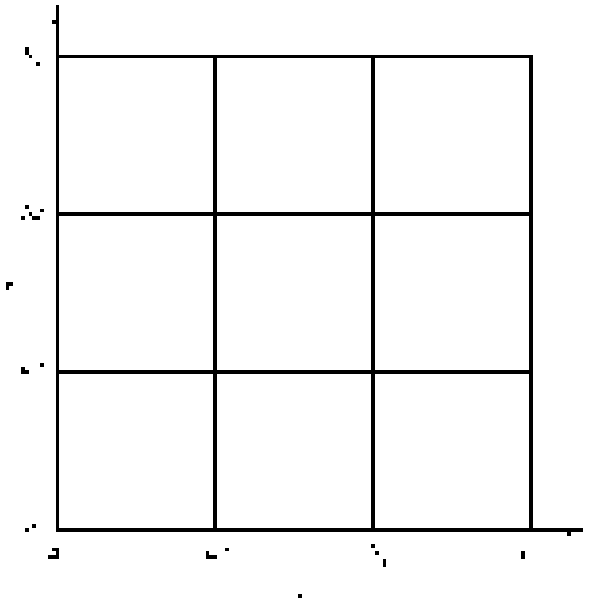

Figure 1.11.

right-open boxes, a given interval I may be expressible as a finite disjoint union of intervals $J_1, \ldots, J_n$, so that on the extension to be well defined we must have $F(I) = \sum_{j=1}^{n} F(J_j)$, but this follows from (2.1).

(vi) *The set function μ is countably subadditive on $\mathcal{S}_1(\mathbb{R}^1)$.*

For each number $\varepsilon > 0$ define $F_\varepsilon : \mathbb{R}^1 \to \mathbb{R}^1$ by $F_\varepsilon(x) = F(x) + \varepsilon x$. Then F_ε is continuous and strictly increasing, so $F_\varepsilon(\mathbb{R}^1) = \mathbb{R}^1$, and the associated function μ_ε on $\mathcal{S}_1(\mathbb{R}^1)$ satisfies all the hypotheses of Theorem 1.4.3.

Now assume $\mu(\mathbb{R}^1) = \infty$. Then $\mathbb{R}^1$ can be written as $\bigcup_{j=-\infty}^{\infty} I_j$, where $I_j = [j, j+1)$, and the finite-valued set function μ_j on $\mathcal{S}_1(\mathbb{R}^1)$ defined by $\mu_j(E) = \mu(E \cap I_j)$ is countably additive on $\mathcal{S}_1(\mathbb{R}^1)$. Hence $\mu(E) = \sum_{j=-\infty}^{\infty} \mu_j(E)$ is countably additive on $\mathcal{S}_1(\mathbb{R}^1)$. This completes the proof of Theorem 2.4.3 in the general case.

2.4.9 *Theorem.* Let F be a distribution function on $\mathbb{R}^1$, and let $\mu_F([a,b]) = F(b) - F(a)$, $a, b \in \mathbb{R}$, $a \le b$. There is a unique extension of μ_F to a Lebesgue–Stieltjes measure on $\mathcal{S}_1$.

Proof. Repeat the proof of 1.4.4, with appropriate minor modifications. For example, in considering a set $E \in \mathcal{S}_1(\mathbb{R}^1)$ the role of finite disjoint unions of right-semiclosed intervals in $\mathbb{R}^1$, we take sequences $\{I_j\}$ of

The one-to-one measure preserving correspondence is given by

$$[a,b) \leftrightarrow (a,b] \qquad a,b \in \mathbb{R}^n$$
$$\text{or if} \qquad b_i \in \mathbb{R}^n \text{ for at least one component of } b \text{ is } -\infty,$$

then if the intervals $[(a_1, \ldots, a_n), (b_1, \ldots, b_n))$ and $(a_1, \ldots, a_n)$ are such that $b_i = \infty$, the corresponding intervals $(a, b]$ are (a_1, ∞). The remainder of the proof is left to the reader. ∎

1.4.20 *Examples.* (a) Let $F_1, F_2, \ldots, F_n$ be distribution functions on $\mathbb{R}$ and define $F(x_1, \ldots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n)$. Then F is a distribution function on $\mathbb{R}^n$ since

$$F((a,b]) = \prod_{i=1}^{n} [F_i(b_i) - F_i(a_i)].$$

In particular if $F_i(x) = x$ $(i = 1, \ldots, n)$ then each F_i corresponds to Lebesgue measure on $\mathcal{B}(\mathbb{R})$. In this case we have $F(x_1, \ldots, x_n) = x_1 x_2 \cdots x_n$ and $m(A) = F(m_1[A], \ldots, m_n[A])$. Thus the measure of any rectangular set is its volume; it is called Lebesgue measure on $\mathcal{B}(\mathbb{R}^n)$. Once more, the completion of $\mathcal{B}(\mathbb{R}^n)$ relative to Lebesgue measure is called the class of Lebesgue measurable sets in $\mathbb{R}^n$, written $\mathcal{B}^*(\mathbb{R}^n)$.

(b) Let f be a nonnegative function from $\mathbb{R}^n$ to $\mathbb{R}$ such that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\,dx_1 \cdots dx_n < \infty$$

(For now, we assume the integration is in the Riemann sense.) Define

$$F(x) = \int_{(-\infty,\ldots,-\infty)}^{x} f(t)\,dt$$

that is,

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\,dt_1 \cdots dt_n.$$

Then

$$\Delta_{a_1,\ldots,a_n} F(x_1, \ldots, x_n) = \int_{a_1}^{x_1} \cdots \int_{a_n}^{x_n} f(t_1, \ldots, t_n)\,dt_1 \cdots dt_n$$

and we find by repeating this computation that

$$F(a,b] = \int_{a_p}^{b_p} \cdots \int_{a_1}^{b_1} f(x_1,\ldots,x_p)\,dx_1\cdots dx_p.$$

Thus F is a distribution function. If μ is the Lebesgue–Stieltjes measure determined by F, we have

$$\mu(a,b] = \int_{a_1,b_1} f(x)\,dx.$$

We have seen that if F is a distribution function on $\mathbb{R}^p$, there is a unique Lebesgue–Stieltjes measure determined by $\mu((a,b]) = F(a,b]$, $a < b$. Also, if μ is a finite measure on $\mathcal{B}(\mathbb{R}^p)$ and $F(x) = \mu(-\infty, x_1] \times \cdots \times (-\infty, x_p]$, then F is a distribution function on $\mathbb{R}^p$ and $\mu(a,b] = F(a,b]$, $a < b$. It is possible to associate a distribution function with every Lebesgue–Stieltjes measure on $\mathcal{B}^p$, and thus establish a one to one correspondence between Lebesgue–Stieltjes measures and distribution functions. Uniqueness of distribution functions with the same increments $F(a,b] = F(b) - F(a)$, $a < b$, are identified. The result will not be needed, and the details are quite tedious and will be omitted.

The following result shows that under appropriate conditions, a Borel set can be approximated from below by a compact set and from above by an open set.

(7.2) Theorem. If μ is a σ-finite measure on $(\mathbb{R}^p, \mathcal{B}^p)$, then for each $B \in \mathcal{B}(\mathbb{R}^p)$,

(a) $\mu(B) = \sup\{\mu(K) : K \subset B, K \text{ compact}\}$
if μ is also a Lebesgue–Stieltjes measure then
(b) $\mu(B) = \inf\{\mu(V) : V \supset B, V \text{ open}\}$.

(c) There is an example of a σ-finite measure on $(\mathbb{R}^1, \mathcal{B}^1)$ that is not a Lebesgue–Stieltjes measure and for which (b) fails.

Proof.

(a) First assume that μ is finite. Let $\mathcal{D}$ be the class of subsets of $\mathbb{R}^p$ having the desired property. We show that $\mathcal{D}$ is a monotone class. Indeed, let $B_n = \bigcup_1^n B_i$, $B_i \in \mathcal{D}$. Then B_n has a compact subset K_n with $\mu(B_n) < \mu(K_n) + \varepsilon/2^n$. If necessary by replacing K_n by $\bigcup_{i=1}^n K_i$ we may assume the K_n form an increasing sequence. Then $\mu(B) = \lim_{n\to\infty} \mu(B_n) < \lim_{n\to\infty} \mu(K_n) + \varepsilon$, so that

$$\mu(B) = \sup\{\mu(K) : K \text{ a compact subset of } B\},$$

and $B \in \mathcal{L}$. If $B_0 \subset \mathbb{R}^n$, $\lambda_n B_0 = B$, set K_1 be a compact subset of B_0 such that $m(B_0) \geq m(K_1) + \varepsilon 2^{-1}$, and set $K = \bigcup_{j=1}^{\infty} K_j$. Then

$$\mu(B) - \mu(K) = m(B) - m(K) \leq m\left(\bigcup_{j=1}^{\infty} \lambda_n B_0 - K_j\right) = \sum_{j=1}^{\infty} m(\lambda_n B_0 - K_j) < \varepsilon,$$

thus $B \in \mathcal{L}$. Therefore $\mathcal{L}$ is a monotone class containing all finite disjoint unions of rectangles (a rectangle is a Cartesian product of intervals; an interval is the limit of an increasing sequence of compact intervals). Hence $\mathcal{L}$ contains all Borel sets.

If μ is σ-finite, each $B \in \mathcal{L}$ with $\mu(B) < \infty$ is the limit of an increasing sequence of sets B_j of finite measure. Each B_j can be approximated from within by compact sets, so by the previous argument m is the measure given by $\mu(A) = \mu(A \cap B_j)$ for $A = \mathcal{L}(\mathbb{R}^n)$, and the preceding argument that $\mathcal{L}$ is closed under limits of increasing sequences shows that $\mathcal{L} = \mathcal{L}$.

(b) We have $\mu(B) \leq \inf\{\mu(W) : W \supset B, \ W \text{ open}\}$
$= \inf\{\mu(W) : W \supset B, \ W - B \text{ compact}\}$

If μ is finite, this equals $\mu(B)$ by (a), applied to $W - B$, and the result follows.

Now assume μ is an arbitrary Borel measure, and write $\mathbb{R}^n = \bigcup_{j=1}^{\infty} P_j$, where the P_j are disjoint bounded sets. Then $P_j \subset U_j$ for some bounded open set U_j. The measure $\mu_j(A) = \mu(A \cap U_j)$, $A \in \mathcal{L}(\mathbb{R}^n)$, is finite, hence if $B \subset P_j$ is a Borel subset of P_j and $\varepsilon > 0$, there is an open set $W_j \supset B$ such that $\mu_j(W_j - B) < \varepsilon 2^{-j}$. Now $W = \bigcup_{j=1}^{\infty} W_j$ is open and $W - B = \bigcup_{j=1}^{\infty} (W_j - B)$, hence $\mu(W - B) \leq \sum_{j=1}^{\infty} \mu(W_j - B) < \varepsilon$. Now any $B \in \mathcal{L}(\mathbb{R}^n)$ is a countable union of sets $B \cap P_j$, each of which is a Borel set plus a set of measure zero. Finally, if μ is σ-finite and $B \in \mathcal{L}$ with $\mu(B) = \infty$, let U_j be an open set with $U_j \supset B \cap P_j$ and $\mu(U_j) \geq \mu(B \cap P_j) + 2^j$. Then $U = \bigcup_{j=1}^{\infty} U_j$ is open, $U \supset B$, and $\mu(U) = \sum_{j=1}^{\infty} \mu(U_j) \geq \mu(B) + 2^j$.

(c) Consider the measure μ on $\mathbb{R}^1$ as in Section 1.5, and let μ be constant on $[j, j+1)$ for $j = 0, 1, 2, \ldots$, and take $\mu(\{j\}) = -j$ for all j. Since $\mathbb{R} = \bigcup_{j=-\infty}^{\infty} [j, j+1)$ and $\mu([j, j+1)) = 0$, μ is σ-finite since

$$\mu([0, k]) = \sum_{j=0}^{k-1} j = -\infty,$$

μ is not a Borel measure because $\mu(\{j\}) = -j$ is not ≥ 0. Now $\mu(U) = 0$ for all U an open set containing 0, we have

$$\mu(U) = \mu([-1, 1)) \qquad \text{for some } j$$
$$= \sum_{j=-1}^{1} j \qquad \text{for some } j$$
$$= -\infty$$

This fifty cents. [Another example. Let $p(x)$ be the number of rational points in $[0,x]$.]

Problems

1. Let F be the cantor function on $\mathbb{R}$ given by $F(x)=0$ if $x\le 0$; $F(x)=1$ if $x\ge 1$; $F(x)=\sum_{k=1}^{n} a_k 2^{-k}$ if $x=\sum_{k=1}^{n} a_k 3^{-k}$, $a_k=0$ or 2, $a_n=2$; and $F(x)=\sup\{F(y): y<x,\ y=\sum_{k=1}^{n} a_k 3^{-k}, a_k=0 \text{ or } 2, a_n=2\}$ if $0<x<1$. Let μ be the Lebesgue–Stieltjes measure corresponding to F, and compute the measure of each of the following sets.
   (a) $\{\frac{1}{3}\}$. (b) $\{\frac{1}{4}\}$.
   (c) $[0,\frac{1}{3}]$. (d) $[\frac{1}{3},\frac{2}{3}]$.
   (e) $\mathbb{Q}\cap[0,1]$. (f) $\mathbb{Q}^c\cap[0,1]$.
   (g) $C\cap[0,1]$, C the Cantor set.

2. Let μ be a Lebesgue–Stieltjes measure on $\mathbb{R}$ corresponding to a continuous distribution function.
   (a) If A is a countable subset of $\mathbb{R}$, show that $\mu(A)=0$.
   (b) If $\mu(A)=0$, must A include an open interval?
   (c) If $\mu(A)>0$ and $\mu(B)=0$, must $A\setminus B$ be dense in A?
   (d) Do the answers to (b) and (c) change if μ is not assumed to be Lebesgue measure?

3. Let Ω be a bounded open set in $\mathbb{R}^n$, and let $\partial\Omega=\{x\in\mathbb{R}^n: x\in\overline{\Omega}\setminus\Omega\}$. Show that $\partial\Omega$ is a Borel set and $\partial\Omega=\{x\in\overline{\Omega}: x\notin\Omega\}$ is a Borel set. [Use the analogous principle.]

4. Show that if $B\in\mathcal{B}(\mathbb{R}^n)$, $\mu(B)<\infty$, $\mathbb{R}^n=\bigcup_{k=1}^{\infty}B_k$, $B_k\in\mathcal{B}(\mathbb{R}^n)$, and $\mu(B_k)=\mu(B)$, where μ is Lebesgue measure. [Thus Lebesgue measure is translation invariant. (The general case can be reduced to consideration of the unit cube, then to cubes with rational vertices, and finally to the countable case.)]

5. Let μ be a Lebesgue–Stieltjes measure on $\mathcal{B}(\mathbb{R}^n)$ such that $\mu(x+B)=\mu(B)$ for all $x\in\mathbb{R}^n$ and for all Borel sets $B\subset\mathbb{R}^n$. In other words, μ is translation invariant on Borel sets. Show that μ is a constant times Lebesgue measure.

6. (a) Show that a non-Lebesgue measurable set $C\subset[0,1]$ and numbers r and s can be found such that $C+r$ and $C+s$ are disjoint, where $C+r=\{x+r: x\in C\}$. [Consider an enumeration of each of the distinct equivalence classes $E_\alpha=\{x: x-y=\text{rational}\}$ formed using $A=[0,1]$. Establish the contradiction.]
   (b) Let C and r and s be as in (a). Let μ be a Lebesgue measure on $\mathcal{B}(\mathbb{R})$ such that $\mu(x+B)=\mu(B)$ for all $x\in\mathbb{R}$ and for all $B\in\mathcal{B}(\mathbb{R})$. Show that $\mu(C)=0$. [Use the relation $\mu(C+r)=\mu(C+s)$.] Conclude that C cannot be Lebesgue measurable.

The only properties of Lebesgue measure needed in this problem are countable subadditivity and finiteness on bounded intervals. The idea is to mimic the Banach-Tarski nonmeasurable translation-invariant subset of $[0,1]$ by choosing $A \subset [0,1]$ so that $\lambda^*(A \cap I) > 0$ for each bounded interval I.

7. *The Cantor function.* Let C be the middle third of the interval $[0,1]$, that is, $F_1 = [\frac{1}{3}, \frac{2}{3}]$, and let $F_2 = [\frac{1}{9}, \frac{2}{9}] \cup [\frac{7}{9}, \frac{8}{9}]$. The set C is obtained in the first step from $[0,1]$ by removing the open middle third of the interval that remains. In the second step C is obtained from $[0,1] - F_1$ by removing the middle thirds of the two intervals that remain. Thus $C = [0,1] - (F_1 \cup F_2 \cup \cdots)$, where F_n consists of 2^{n-1} open intervals each of length 3^{-n}. Each point in C can be written in ternary form using digits 0 and 2 in the first n digits. Continue the construction and let F_n be the union of the middle thirds of the intervals that remain after $F_1, \ldots, F_{n-1}$ are removed. The Cantor ternary set C is defined as $[0,1] - \bigcup_{n=1}^{\infty} F_n$. Each $x \in C$ can be expressed in terms of digits using only digits 0 and 2. We have topological properties of C follow from the definition. For example, each point of C is a limit point of C but contains no interval.

   Show that C is uncountable and has Lebesgue measure 0.

   Comment. In the above construction, we have $m(F_n) = \frac{1}{3}(\frac{2}{3})^{n-1}$, $n = 1, 2, \ldots$, where m is Lebesgue measure. If $0 < \alpha < 1$, the procedure may be altered slightly so that $m(F_n) = \alpha(\frac{2}{3})^{n-1}$. We then obtain a set C_α homeomorphic to C, of measure $1 - \alpha$, and also an example of a set of positive measure

8. Show an example of a function $F : \mathbb{R}^2 \to \mathbb{R}$ such that F is right-continuous and is increasing in each coordinate separately, but F is not a distribution function on $\mathbb{R}^2$.

9. A distribution function on $\mathbb{R}$ is monotone and has only countably many points of discontinuity. Is this also true for a distribution function on $\mathbb{R}^2$? Explain.

10. (a) Let F and G be distribution functions on $\mathbb{R}^1$. If $F[a,b] = G[a,b]$ for all $a < b$, $-\infty < a < b < \infty$, does it follow that F and G differ by a constant?

    (b) Make a distribution function on $\mathbb{R}^2$ be increasing in each coordinate separately?

11. Let $\aleph$ be the cardinality of the reals. Show that there are only $\aleph$ Borel subsets of $\mathbb{R}^1$, but $2^{\aleph}$ Lebesgue measurable sets.

1.8 Integration with respect to an outer measure

If f is a real-valued function on a bounded interval $[a,b]$ of reals, we can talk about the Riemann integral of f, at least if f is piecewise continuous. We are going to develop a much more general integration process, one

that applies to functions from an arbitrary set to the extended reals, provided that certain "measurability" conditions are satisfied.

Preliminary remarks. We may again be used to motivate the concept of measurability. Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space, and that X is a function from Ω to $\mathbb{R}$. Then if the outcome of the experiment corresponds to the point $\omega \in \Omega$, we may associate the number $X(\omega)$ with the experiment. Suppose that we are interested in the probability that X takes a value in a given set A; in other words, we wish to compute $P(X \in A)$, where $A \subseteq \mathbb{R}$. For this to be possible, the set $\{\omega \in \Omega : X(\omega) \in A\}$ must belong to the σ-field $\mathcal{F}$. If $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{F})$ for such intervals A and hence, as we shall see below, for all Borel sets $A \subseteq \mathbb{R}$, then X is a "measurable function" on the sample space, and probabilities of events involving X can be computed. In the language of probability theory, X is a "random variable."

(5.1) Definition and Theorem. Let $f : \Omega_1 \to \Omega_2$ be a function from a set Ω_1 to a set Ω_2, and let $\mathcal{F}_1$ and $\mathcal{F}_2$ be σ-fields of subsets of Ω_1 and Ω_2, respectively. Then the following are equivalent:

(a) $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{F}_2$.

(b) $f^{-1}(A) \in \mathcal{F}_1$ for each A in some collection $\mathcal{A}$ of subsets of Ω_2 such that $\sigma(\mathcal{A}) = \mathcal{F}_2$.

(c) $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a semiring of subsets of Ω_2 such that $\sigma(\mathcal{A}) = \mathcal{F}_2$.

(d) $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a field of subsets of Ω_2 such that $\sigma(\mathcal{A}) = \mathcal{F}_2$.

(e) $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a collection of subsets of Ω_2 such that $\mathcal{F}_2$ is the smallest monotone class containing $\mathcal{A}$.

(f) $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a collection of subsets of Ω_2 such that $\mathcal{F}_2$ is the smallest σ-field containing $\mathcal{A}$.

If these conditions hold, then f is said to be $(\mathcal{F}_1, \mathcal{F}_2)$-measurable, or simply measurable if $\mathcal{F}_1$ and $\mathcal{F}_2$ are clear from the context.

Proof. (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (e) $\Rightarrow$ (f) are trivial. (f) $\Rightarrow$ (a): Let $\mathcal{M} = \{A \subseteq \Omega_2 : f^{-1}(A) \in \mathcal{F}_1\}$. Then $\mathcal{M}$ is a σ-field and $\mathcal{A} \subseteq \mathcal{M}$, so $\mathcal{F}_2 = \sigma(\mathcal{A}) \subseteq \mathcal{M}$. Thus (a) holds. ∎

It is sufficient that $f^{-1}(A) \in \mathcal{F}_1$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a collection of subsets of Ω_2 such that the minimal σ-field over $\mathcal{A}$ is $\mathcal{F}_2$. For $\{A \subseteq \Omega_2 : f^{-1}(A) \in \mathcal{F}_1\}$ is a σ-field that contains all sets of $\mathcal{A}$, hence coincides with $\mathcal{F}_2$. This is another application of the good sets principle.

The notation $f : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ will mean that $f : \Omega_1 \to \Omega_2$ is measurable relative to $\mathcal{F}_1$ and $\mathcal{F}_2$.

If $\mathcal{F}$ is a σ-field of subsets of Ω, then $(\Omega, \mathcal{F})$ is sometimes called a measurable space and the sets in $\mathcal{F}$ are sometimes called measurable sets.

Notice that measurability of f does not imply that $f(A) \in \mathcal{F}_2$ for each $A \in \mathcal{F}_1$. For example, if $\mathcal{F}_2 = \{\Omega_2, \emptyset\}$, then any $f : \Omega_1 \to \Omega_2$ is measurable, regardless of $\mathcal{F}_1$. But if $A \in \mathcal{F}_1$ and $f(A)$ is a nontrivial proper subset of Ω_2, then $f(A) \notin \mathcal{F}_2$. Actually, in measure theory, the inverse image is much more important than the direct image since the basic set operations are preserved by inverse images but not in general by direct images. Specifically, $f^{-1}(\bigcup_i B_i) = \bigcup_i f^{-1}(B_i)$, $f^{-1}(\bigcap_i B_i) = \bigcap_i f^{-1}(B_i)$, $f^{-1}(B^c) = [f^{-1}(B)]^c$, $f^{-1}(\emptyset) = \emptyset$, and $f^{-1}(\Omega_2) = \Omega_1$. But $f(\bigcup_i A_i) = \bigcup_i f(A_i)$, $f(\bigcap_i A_i) \subseteq \bigcap_i f(A_i)$, and the inclusion may be proper. Information, like water, must be poured through f, not f^{-1}, when f is a material function. The two sets are distinct.

If $(\Omega, \mathcal{F})$ is a measurable space and $f : \Omega \to \mathbb{R}^n$ (or $\mathbb{R}^\infty$), it is said to be Borel measurable [or $(\mathcal{F}, \mathcal{B}^n)$ or $\mathcal{B}^\infty$-measurable] if f is measurable relative to the σ-fields $\mathcal{F}$ and $\mathcal{B}^n$ (or $\mathcal{B}^\infty$). If Ω is a Borel set in $\mathbb{R}^n$ [or $\mathbb{R}^\infty$], and we say the term "Borel measurable" we always assume that $\mathcal{F} = \mathcal{B}^n$.

A convenient way to show that f or f^{-1} is Borel measurable, for $\mathbb{R}^n$ to $\mathbb{R}^n$, is to find a class $\mathcal{A}$ of open subsets of $\mathbb{R}^n$ such that $f^{-1}(A)$ is open [or $f(A)$ is open] for each $A \in \mathcal{A}$. Since $\mathcal{B}^n$ is the smallest σ-field containing the open sets, it is the smallest σ-field containing $\mathcal{A}$, and the result follows from (5.1).

If A is a subset of $\mathbb{R}$ that is not a Borel set (Section 1.1, Problems 5 and 6) and f is the indicator of A, that is, $f(x) = 1$ if $x \in A$ and $f(x) = 0$ if $x \notin A$, then f is not Borel measurable, but $[a, \infty)$ if $a \le 1$, $f^{-1}([a, \infty)) = A$ if $0 < a \le 1$, and $\emptyset$ if $a > 1$.

To convince ourselves that $f : \Omega \to \mathbb{R}$ is Borel measurable, it is sufficient to show that $f^{-1}((a, \infty)) \in \mathcal{F}$ for each real a. (Why?) If $\mathcal{C}$ is the class of sets $C \subset \mathbb{R}$ such that $f^{-1}(C) \in \mathcal{F}$, then $\mathcal{C}$ is a σ-algebra. (Why?) Similarly, to show that $f(\omega) \ge 0$ for all ω, it is sufficient to show that $f^{-1}([a, \infty)) = \emptyset$ for all $a < 0$. (Why?)

If $(\Omega, \mathcal{F}, \mu)$ is a measure space and f is a Borel measurable function on $(\Omega, \mathcal{F})$, we will mean that f is Borel measurable on $(\Omega, \mathcal{F})$ and not that f is measurable on $(\Omega, \mathcal{F}, \mu)$.

1.5.2 *Definition.* Let $(\Omega, \mathcal{F})$ be a measurable space, and let $A \in \mathcal{F}$. The indicator function of A, $I_A : \Omega \to \mathbb{R}$, is defined by $I_A(\omega) = 1$ if $\omega \in A$, $I_A(\omega) = 0$ if $\omega \notin A$. A function $\varphi : \Omega \to \mathbb{R}$ is called simple if it is Borel measurable and takes on only finitely many distinct values. Equivalently, φ is simple if it can be written in the form $\sum_{k=1}^{n} a_k I_{A_k}$, where the a_k are distinct real numbers in $\mathbb{R}$ and $A_k = \varphi^{-1}(\{a_k\}) \in \mathcal{F}$. The a_k need not be distinct.

We use the standard arithmetic of $\mathbb{R}$, if $a \in \mathbb{R}$, $-\infty < a < \infty$, $a + \infty = \infty + a = \infty$, $a - \infty = -\infty + a = -\infty$, $a \cdot \infty = \infty \cdot a = \infty$ if $a > 0$, $a \cdot \infty = -\infty \cdot a = -\infty$ if $a < 0$, $0 \cdot \infty = \infty \cdot 0 = 0$, $\infty + \infty = \infty$, $-\infty - \infty = -\infty$, with commutativity of addition and multiplication. It is then easy to check that arithmetic, inequality, and limit operations are well defined, as long as the operations are well defined. In other words we can carry out addition and multiplication of extended real numbers.

Let μ be a measure on $\mathcal{F}$, a sigma-field of subsets of Ω. For $\varphi \ge 0$ a Borel measurable simple function, we are going to define the integral of φ with respect to μ, which will be a number in $[0, \infty]$ which will be denoted by $\int_\Omega \varphi \, d\mu$.

1.5.3 *Definition of the Integral.* First let φ be a simple step function $\varphi = \sum_{k=1}^{n} a_k I_{A_k}$ where the A_k are disjoint sets in $\mathcal{F}$. We define

$$\int_\Omega \varphi \, d\mu = \sum_{k=1}^{n} a_k \mu(A_k)$$

as long as $+\infty$ and $-\infty$ do not both appear in the sum; if they do we say the integral does not exist. Since φ is simple, it can be written in many different representations, say $\sum_{j=1}^{m} b_j I_{B_j}$, but

$$\sum_{k=1}^{n} a_k \mu(A_k) = \sum_{j=1}^{m} b_j \mu(B_j).$$

For example, if $\lambda = \mathcal{L}(n)$, where $\mathcal{L}(n)$ is Lebesgue measure on $\mathbb{R}^n$, then $\nu_\lambda = \mathcal{L}(n)$. The proof is based on the decomposition that

$$A = \bigcup_{r \in \mathbb{Q}} \left(A \cap f^{-1}(-\infty, r) \right),$$

where $\mathbb{Q} = \{r_1, r_2, \ldots\}$. Then

$$\begin{aligned} \sum_{i=1}^{\infty} \nu_\lambda(A_i \cap B_j) &= \sum_{i=1}^{\infty} \lambda \left(A_i \cap f^{-1}(-\infty, r_j) \right) \\ &= \lambda \left(\bigcup_{i=1}^{\infty} A_i \cap f^{-1}(-\infty, r_j) \right) \\ &= \lambda \left(A \cap f^{-1}(-\infty, r_j) \right) \qquad \text{by a symmetrical argument.} \end{aligned}$$

If λ is nonnegative Borel measurable define

$$\int_A \lambda \, d\mu = \sup \left\{ \int_A s \, d\mu : s \text{ simple}, \quad 0 \le s \le \lambda \right\}.$$

This agrees with the previous definition if λ is simple. Furthermore we may have $\int_A \lambda \, d\mu < \infty$ even if λ is not finite-valued.

Notice that according to this definition, the integral of a nonnegative function always exists, it may be $+\infty$.

Finally, if λ is an arbitrary Borel measurable function, let $\lambda^+ = \max(\lambda, 0)$, $\lambda^- = \max(-\lambda, 0)$, that is,

$$\lambda^+(x) = \begin{cases} \lambda(x) & \text{if } \lambda(x) \ge 0 \\ 0 & \text{if } \lambda(x) < 0 \end{cases} \qquad \lambda^-(x) = \begin{cases} 0 & \text{if } \lambda(x) \ge 0 \\ -\lambda(x) & \text{if } \lambda(x) < 0 \end{cases}$$

The function λ^+ is called the positive part of λ, the negative part of λ. We have $\lambda = \lambda^+ - \lambda^-$, $|\lambda| = \lambda^+ + \lambda^-$. Since λ and $|\lambda|$ are Borel measurable, so are λ^+ and λ^-. For example, $\{x : \lambda^+(x) \in B\} = \{x : \lambda(x) \ge 0, \lambda(x) \in B\} \cup \{x : \lambda(x) < 0, 0 \in B\}$. The first set is $\lambda^{-1}(B \cap [0, \infty))$ or $\lambda^{-1}(B)$ according as $0 \in B$ or $0 \notin B$. This is true for each $B \in \mathcal{B}(\mathbb{R})$ and similarly for λ^-. Consequently, if λ_1 and λ_2 are Borel measurable, then $\max(\lambda_1, \lambda_2)$ and $\min(\lambda_1, \lambda_2)$ are Borel measurable, since, for example,

$$\{x : \max(\lambda_1(x), \lambda_2(x)) \le c\} = \{x : \lambda_1(x) \le c\} \cap \{x : \lambda_2(x) \le c\}$$

and $x : \max(\lambda_1(x), \lambda_2(x)) > c\} = \{x : \lambda_1(x) > c\} \cup \{x : \lambda_2(x) > c\}$. In fact if λ is Borel measurable so are $|\lambda|$ and λ^2.

We define

$$\int_E f\,d\mu = \int_E f^+\,d\mu - \int_E f^-\,d\mu \qquad \text{if this is not of the form } \infty - \infty.$$

If it is not clear that the integral does not exist, the function f is said to be μ-integrable (or simply integrable if μ is understood). That $\int_E f\,d\mu$ exists means that, that is, if $\int_E f^+\,d\mu$ and $\int_E f^-\,d\mu$ are both finite.

If $A \in \mathcal{F}$, we define

$$\int_A f\,d\mu = \int_E f I_A\,d\mu.$$

(Use Exercise 1.5.3 to show that fI_A is measurable.) Similar to the last result above, show that $A \mapsto \int_A f\,d\mu$ is Borel measurable.)

If f is a simple function on $\mathbb{R}$ to $\mathbb{R}$ and μ is Lebesgue measure, $\int_E f\,d\mu$ agrees with the Riemann integral. However, the integral of f with respect to Lebesgue measure exists for many functions that are not Riemann integrable, as we shall see in 1.7.

Before examining the properties of the integral, we need to know more about Borel measurable functions. One of the basic reasons why such functions are useful in probability is that a pointwise limit of Borel measurable functions is still Borel measurable.

1.5.4 (Theorem) *If $f_1, f_2, \ldots$ are Borel measurable functions from Ω to $\mathbb{R}$ and $\lim_{n\to\infty} f_n(\omega)$ exists for all $\omega \in \Omega$, then f is Borel measurable.*

(Hint: It is sufficient to show that $\{\omega : \lim_{n\to\infty} f_n(\omega) \le a\} \in \mathcal{F}$ for each real a. We have

$$\begin{aligned}
\{\omega : \lim_{n\to\infty} f_n(\omega) \le a\} &= \left\{\omega : \limsup_{n\to\infty} f_n(\omega) \le a\right\} \\
&= \left\{\omega : f_n(\omega) \text{ is eventually } \le a + \tfrac{1}{r} \text{ for every } r = 1, 2, \ldots\right\} \\
&= \bigcap_{r=1}^{\infty} \left\{\omega : f_n(\omega) \le a + \tfrac{1}{r} \text{ for all but finitely many } n\right\} \\
&= \bigcap_{r=1}^{\infty} \liminf_{n\to\infty} \left\{\omega : f_n(\omega) \le a + \tfrac{1}{r}\right\} \\
&= \bigcap_{r=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{\omega : f_n(\omega) \le a + \tfrac{1}{r}\right\} \in \mathcal{F}
\end{aligned}$$

To show that the class of Borel measurable functions is closed under algebraic operations, we need the following basic approximation theorem.

1.5.5 *Definition.* (a) A nonnegative Borel measurable function h is the limit of an increasing sequence of nonnegative, finite valued, simple functions h_n.

(b) An arbitrary Borel measurable function g is the limit of a sequence of finite-valued simple functions g_n such that $|g_n| \le |g|$ for all n.

Proof. (a) Define

$$h_n(\omega) = \frac{k-1}{2^n} \quad \text{if} \quad \frac{k-1}{2^n} \le h(\omega) < \frac{k}{2^n}, \qquad k = 1, 2, \ldots, n2^n,$$

and let $h_n(\omega) = n$ if $h(\omega) \ge n$. [In words, $h_n(\omega) = (k-1)/2^n$ if $h(\omega)$ lies between the $(k-1)/2^n$ and $k/2^n$, $k = 1, 2, \ldots, n2^n$; $h_n(\omega) = n$ if $h(\omega) \ge n$.] The h_n have the desired properties (Problem 11).

(b) Let g_n and h_n be nonnegative, finite valued, simple functions with $g_n \uparrow g^+$ and $h_n \uparrow g^-$; take $f_n = g_n - h_n$. ∎

1.5.6 *Theorem.* If h_1 and h_2 are Borel measurable functions from Ω to $\mathbb{R}$, so are $h_1 + h_2$, $h_1 h_2$, $|h_1|$, $\max(h_1, h_2)$, and $\min(h_1, h_2)$ [assuming these are well-defined; in other words, $h_1(\omega) + h_2(\omega)$ is not of the form $+\infty - \infty$ and $h_1(\omega)/h_2(\omega)$ is not of the form $\pm\infty/\pm\infty$ or $g/0$].

Proof. As in 1.5.5, let $g_n \uparrow h_1$ and $h_n \uparrow h_2$ be finite-valued simple functions with $g_1 \le g_2 \le \cdots$ and $h_1 \le h_2 \le \cdots$. Then $g_n + h_n \uparrow h_1 + h_2$, $g_n h_n \uparrow h_1 h_2$, $|g_n| \uparrow |h_1|$, and

$$\frac{g_n}{g_n + (1/n)\mathbf{1}_{\{g_n = 0\}}} \to \frac{h_1}{h_2}$$

Since

$$\max(g_n, h_n) = \frac{1}{2}(g_n + h_n + |g_n - h_n|) \to \frac{1}{2}\left(h_1 + h_2 + |h_1 - h_2|\right) = \max(h_1, h_2)$$

we conclude the result follows immediately. ∎

Whereas going from 1.5.4 to 1.5.6 in part (a) of 1.5.4 in Borel measurable functions from Ω to $\mathbb{R}^n$, to do this, we need the following useful result.

1.5.7 *Lemma.* A composition of measurable functions is measurable: If $g: (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ and $h: (\Omega_2, \mathcal{F}_2) \to (\Omega_3, \mathcal{F}_3)$, then $h \circ g: (\Omega_1, \mathcal{F}_1) \to (\Omega_3, \mathcal{F}_3)$.

Proof. If $B \in \mathcal{B}_Y$, then $(f \circ g)^{-1}(B) = g^{-1}(f^{-1}(B)) \in \mathcal{B}_Z$. ∎

Since some books contain the statement "A composition of measurable functions need not be measurable," some explanation is called for. If $f: X \to Y$, some authors call f "measurable" if the preimage of a Borel set is a Lebesgue measurable set. We shall call such a function Lebesgue measurable. Note that every Borel measurable function is Lebesgue measurable, but not conversely. Consider the indicator of a Lebesgue measurable set that is not a Borel set (see Section 1.4). Problem 1.11 (c) says that if f and g are Lebesgue measurable, the composition $g \circ f$ need not be Lebesgue measurable. But if f is the Borel measurable function $\overline{\chi}_C$ and g is Lebesgue measurable, then $g \circ f = \overline{\chi}_C$ has the property that $(g \circ f)^{-1}(B) \in \mathcal{L}_X$ for every $B \in \mathcal{B}_{\mathbb{R}}$ (see Problem 1.11 (b)). It is known that $g \circ f$ is Lebesgue measurable only when $\lambda^*(C) = 0$ or $\lambda^*(C^c) = 0$. For an explicit example, see Royden [1968, p. 70]. If $g: \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable and $f: \mathbb{R} \to \mathbb{R}$ is continuous, then $g \circ f$ is Lebesgue measurable. Indeed, if $B \in \mathcal{B}_{\mathbb{R}}$, then $(g \circ f)^{-1}(B) = f^{-1}(g^{-1}(B)) \in \mathcal{B}_{\mathbb{R}}$ because $g^{-1}(B) \in \mathcal{L}_{\mathbb{R}}$ by Theorem 1.5.7 and $f^{-1}(g^{-1}(B))$ is Lebesgue measurable; similarly, if f is Borel measurable and g is Borel measurable, then $g \circ f$ is Lebesgue measurable.

It is rarely necessary to replace Borel measurability of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ by the slightly more general concept of Lebesgue measurability, in this book. The only instance is in 1.7. This is explained by the fact that we are developing tools for a general treatment of measures on an arbitrary metric space (X, ρ). Then there is no procedure to map Lebesgue measurable functions on $\Omega = [0,1]$, $\mathcal{F} = \mathcal{L}_{[0,1]}$, $P = \lambda$ onto X.

We may now state some of the $\mathcal{A}_1, \mathcal{A}_2, \ldots$ pre-Borel measurable functions from Ω to $\mathbb{R}^1$ and $\mathcal{A}_1$ consists of all intervals in $\mathbb{R}$. Then it is Borel measurable. Furthermore, if $\mathcal{A}_1$ and $\mathcal{A}_2$ are Borel measurable functions from Ω to $\mathbb{R}^1$, so are $\mathcal{A}_1 + \mathcal{A}_2$ and $\mathcal{A}_1 \mathcal{A}_2$. Assuming these are well-defined. The function $\mathcal{A}(\omega) = (\mathcal{A}_1(\omega), \ldots, \mathcal{A}_n(\omega))$ describes a map from Ω to $\mathbb{R}^n$. Borel measurability of $\mathcal{A}$ is equivalent to Borel measurability of the component functions $\mathcal{A}_i$.

1.5.8 Theorem. Let $\mathcal{A}: \Omega \to \mathbb{R}^n$. If p_i is the projection map of $\mathbb{R}^n$ onto $\mathbb{R}$, taking $(x_1, \ldots, x_n)$ to x_i, and $\mathcal{A}_i = p_i \circ \mathcal{A}$, $i = 1, \ldots, n$, then $\mathcal{A}$ is Borel measurable iff $\mathcal{A}_i$ is Borel measurable for all $i = 1, \ldots, n$.

Proof. Assume $\mathcal{A}$ Borel measurable. Since

$$p_i^{-1}(a_i, b_i) = \{x \in \mathbb{R}^n : a_i < x_i < b_i\} = \mathbb{R} \times \cdots \times \mathbb{R} \times (a_i, b_i) \times \mathbb{R} \times \cdots \times \mathbb{R} \quad (i \text{ times}),$$

which is an interval in $\mathbb{R}^n$, p_i is Borel measurable. Thus

$$\mathcal{A}_i^{-1}(B) = (p_i \circ \mathcal{A})^{-1}(B) = \mathcal{A}^{-1}(p_i^{-1}(B)) \in \mathcal{F} \quad \text{for all } B \in \mathcal{B}_{\mathbb{R}},$$

and therefore by 1.5.6, $\mathcal{A}_i: (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Conversely, assume each f_n to be Borel measurable. Then

$$\lambda^m\{x \in \mathbb{R}^m : a_i \le x_i < b_i,\ i = 1, \dots, m\} = \sum_{n=1}^{\infty} \{x \in \Omega : a_i \le f_i(x) < b_i,\ i = 1, \dots, m\},$$

and the result follows. $\square$

We now proceed to define the integral of measurable functions. Throughout the following discussion, $(\Omega, \mathcal{A}, \mu)$ denotes a fixed measure space.

5.3.9 *Definition.* (a) If $\int_\Omega s\,d\mu$ exists and $c \in \mathbb{R}$, then $\int_\Omega cs\,d\mu$ exists and equals $c\int_\Omega s\,d\mu$.

(b) If $\int_\Omega s\,d\mu < \infty$ for all s_i then $\int_\Omega \sum_i s_i\,d\mu = \sum_i \int_\Omega s_i\,d\mu$ in the sense that if $\int_\Omega s_i\,d\mu$ exists and is finite for $i = 1, \dots, n$, then $\int_\Omega \sum_{i=1}^n s_i\,d\mu$ exists and $\int_\Omega \sum_{i=1}^n s_i\,d\mu = \sum_{i=1}^n \int_\Omega s_i\,d\mu$. Thus if each integral exists, $\int_\Omega (s_1 + s_2)\,d\mu = \int_\Omega s_1\,d\mu + \int_\Omega s_2\,d\mu$, whether or not the integrals are finite.

(c) If $\int_\Omega s\,d\mu$ exists, then $\int_A s\,d\mu = \int_\Omega s\chi_A\,d\mu$.

(d) If $s = 0$ a.e. $[\mu]$ on $A \in \mathcal{A}$, then $\int_A s\,d\mu = 0$ for $\int_\Omega s\,d\mu$. If $s \le t$ a.e. $[\mu]$ on A, then $\int_A s\,d\mu \le \int_A t\,d\mu$.

(e) If $\int_\Omega s\,d\mu$ exists and $\int_\Omega s\,d\mu \ge 0$ for each $A \in \mathcal{A}$, if $\int_\Omega s\,d\mu$ is finite for some A, then $\int_A s\,d\mu$ is also finite for each $A \in \mathcal{A}$.

Proof. (a) is immediate, and the finite case of (b) is similar. (c) is meaningful when $s\chi_A$ is integrable, and if $s = \sum_{i=1}^n a_i \chi_{A_i}$, then

$$\int_A s\,d\mu = \sup\left\{\int_A t\,d\mu : t \text{ is a measurable simple function with } t \le s\right\}$$

$$= \sup\left\{\int_\Omega \sum_{i=1}^n b_i \chi_{A \cap A_i}\,d\mu : 0 \le b_i \le a_i \text{ for } i = 1, \dots, n \text{ and } \sum_{i=1}^n b_i \chi_{A \cap A_i} \text{ is simple}\right\} = \int_\Omega s\chi_A\,d\mu.$$

In general, if $s = s^+ - s^-$ and $s\chi_A = s^+\chi_A - s^-\chi_A$; hence $\int_A s\,d\mu = \int_\Omega s^+\chi_A\,d\mu - \int_\Omega s^-\chi_A\,d\mu$ by what we have just proved, so that $\int_\Omega s\chi_A\,d\mu = \int_A s\,d\mu$. If $s = t$ a.e., then

$$(s - t)^+ = 0 \text{ a.e.}, \qquad (s - t)^- = 0 \text{ a.e.},$$

so

$$\int_A s\,d\mu = \int_A s^+\,d\mu - \int_A s^-\,d\mu = \int_A t^+\,d\mu - \int_A t^-\,d\mu = \int_A t\,d\mu.$$

(d) If s and t are nonnegative and $0 \le s \le t$, then $0 \le \int_\Omega s\,d\mu \le \int_\Omega t\,d\mu$. Hence $\int_A s\,d\mu \le \int_A t\,d\mu$. In general, $s \le t$ implies $s^+ \le t^+$ and $s^- \ge t^-$. If

$\int_\Omega h\,d\mu < \infty$, we have $\int_\Omega u^+\,d\mu \le \int_\Omega u\,d\mu + \varepsilon$. Hence $\int_\Omega h\,d\mu$ exists and equals

$$\int_\Omega h^+\,d\mu - \int_\Omega h^-\,d\mu \le \int_\Omega u^+\,d\mu - \int_\Omega u^-\,d\mu + \varepsilon = \int_\Omega u\,d\mu + \varepsilon \le \int_\Omega h\,d\mu + 2\varepsilon.$$

Since ε is arbitrary, $\int_\Omega h\,d\mu \le \int_\Omega u\,d\mu$. The reverse inequality follows similarly.

(iii) We have $|h| \le h \le |h|$ on Ω. Hence by (ii) and (i), $-\int_\Omega |h|\,d\mu \le \int_\Omega h\,d\mu \le \int_\Omega |h|\,d\mu$, and the result follows. Note that $|h|$ is Borel measurable by 1.5.5 since $|h| = h^+ + h^-$.

(iv) If $0 \le h \le k$, then $\int_\Omega h\,d\mu \le \int_\Omega k\,d\mu$ by (ii). Hence

$$\int_\Omega h\,d\mu = \sup\left\{ \int_\Omega s\,d\mu : s \text{ is simple and } 0 \le s \le h \right\}.$$

If $0 \le s \le h$ is simple, then $s = \sum_{i=1}^n a_i \chi_{A_i}$ and $\int_\Omega s\,d\mu = \sum_{i=1}^n a_i \mu(A_i) \le \int_\Omega k\,d\mu$. Take the sup over s to obtain $\int_\Omega h\,d\mu \le \int_\Omega k\,d\mu$. If $g \le k$ is simple,

(v) This follows from (iv) and the fact that $\mu(A \cap B) \le \mu(A)$, $\int_B g\,d\mu = \int_\Omega g\chi_B\,d\mu \le \int_\Omega g\chi_A\,d\mu$.

Problems

1. Show that the functions proposed in the proof of 1.5.3(a) have the desired properties. Show also that if λ is bounded, the approximating sequence converges uniformly on Ω.

2. Let f and g be extended real-valued Borel measurable functions on $(\Omega, \mathcal{F})$, and define

$$h(\omega) = f^2(\omega) \quad \text{if} \quad \omega \in A_1,$$
$$= g(\omega) \quad \text{if} \quad \omega \in A_2,$$

where A_1 is a set in $\mathcal{F}$. Show that h is Borel measurable.

3. If $f_1, f_2, \ldots$ are extended real-valued Borel measurable functions on $(\Omega, \mathcal{F})$, $a_n = 1, 2, \ldots$, show that $\sup_n f_n$ and $\inf_n f_n$ are Borel measurable. Hence conclude that $\limsup_{n\to\infty} f_n$ and $\liminf_{n\to\infty} f_n$ are Borel measurable.

4. Let $(\Omega, \mathcal{F}, \mu)$ be a complete measure space, let $f : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, and let $g : \Omega \to \Omega'$, $g = f$ except on a subset of a set $A \in \mathcal{F}$ with $\mu(A) = 0$. Show that g is measurable relative to $\mathcal{F}$ and $\mathcal{F}'$.

5. Let f be a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, not necessarily Borel measurable. Show that if f is almost continuous, i.e., if $f = g$ a.e. F_μ on $\mathbb{R}^n$ where

minimal closed subset of $\mathbb{R}^k$ and hence is a Borel set. Does this result hold in spaces more general than the Euclidean space $\mathbb{R}^k$?

(b*) Show that there is no function from $\mathbb{R}$ to $\mathbb{R}$ whose discontinuity set is the rationals. (*Hint*: 1.4.5 was constructed a distribution function whose discontinuity set was the rationals.)

16 How many Borel measurable bijections are there from $\mathbb{R}^n$ to $\mathbb{R}^n$?

17 We have seen that a pointwise limit of measurable functions is measurable. We may also show that under certain conditions a pointwise limit of measures is a measure. The following result, known as Steinhaus' theorem, will be needed in the problems. If $\{a_{ij}\}$ is a double sequence of real numbers satisfying

(i) $\sum_{j=1}^{\infty} |a_{ij}| < \infty$ for all i,
(ii) $\sum_{i=1}^{\infty} |a_{ij}| = \infty$ for all j, and
(iii) $a_{ij} \geq 0$ if $a_{ij} \neq 0$,

then there is a sequence $\{\lambda_i\}$ with $\lambda_i = 0$ or 1 for all i such that $\lambda_i = \sum_{i=1}^{\infty} \lambda_i a_{ij}$ does not converge to a finite or infinite limit.

To prove this, choose positive integers n_1 and k_1 arbitrarily; having chosen $n_1 < n_2 < \cdots < n_{k-1}$ and $k_1, \ldots, k_{n-1}$, choose $n_k > n_{k-1}$ such that $\sum_{i=n_k}^{\infty} |a_{ik_j}| < \frac{1}{2}$; this is possible by (ii). Now choose $k_n > k_{n-1}$ such that $\sum_{i=1}^{n_k} |a_{ik_n}| < \frac{1}{2}$; this is possible by (ii). Set $\lambda_i = 0$ if $i \neq k_j$ for all j, and $\lambda_{k_j} = 1$ if $\lambda_{k_{j-1}} = 1$. We may write $\lambda_{k_j} = \lambda_{k_1} + \lambda_{k_2} + \cdots + \lambda_{k_j}$, where λ_{k_1} is the sum of a_{ik_1}'s. But $k_1 \leq k_2 \leq k_3$ corresponds to $\lambda_{k_1} = 1$, and k_2 to $\lambda_{k_2} = 0$. If $\lambda_{k_3} = 1$, then $k_3 \leq k_4$, because $\lambda_{k_4} = 1$ or 0. Therefore, $\lambda_{k_j} = \sum_{i=1}^{k_j} a_{ik_j}$ is dominated by (i).

$$\lambda_{k_j} = \sum_{i<k_1} a_{ik_j} + \sum_{k_1 \leq i < k_2} a_{ik_j} + \cdots + \sum_{k_{j-1} \leq i < k_j} a_{ik_j}$$

Thus $\lambda_{k_j} < \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^j} < 1$, so $\{\lambda_i\}$ cannot converge.

18 (*Vitali–Hahn–Saks Theorem*. Let $(\Omega, \mathcal{F})$ be a measurable space and let P_n, $n = 1, 2, \ldots$, be probability measures on $\mathcal{F}$. If $P_n(A) \to P(A)$ for all $A \in \mathcal{F}$, then P is a probability measure on $\mathcal{F}$ (that is, P is countably additive) and $\sup_n P_n(B) \downarrow 0$ if $B \downarrow \emptyset$. (Let $A_1, A_2, \ldots$ be disjoint sets in $\mathcal{F}$, let $\varepsilon > 0$, and choose n_0 so large that $\sum_{n=n_0+1}^{\infty} P_n(\bigcup_{i=1}^{\infty} A_i) < \varepsilon$. Then choose k_0 so large that $\sum_{i=k_0+1}^{\infty} P_n(A_i) < \varepsilon$ for $n = 1, \ldots, n_0$. Now $\sum_{i=1}^{\infty} P_n(A_i) = \sum_{i=1}^{k_0} P_n(A_i) + \sum_{i=k_0+1}^{\infty} P_n(A_i) < \sum_{i=1}^{k_0} P_n(A_i) + \varepsilon$ for all n. Let $n \to \infty$ and apply Steinhaus' theorem.)

(b) Extend the Vitali–Hahn–Saks theorem to the case where the P_n are not necessarily probability measures, but $P_n(\Omega) \le c < \infty$ for all n. For further extensions see Dunford and Schwartz (1958).

1.6 Basic Integration Theorems

We are now ready to present the main properties of the integral. The results in this section will be used many times in the text. As always, $(\Omega, \mathcal{F}, \mu)$ is a fixed measure space and all functions to be considered map Ω into $\mathbb{R}$.

1.6.1 Theorem. Let f be a Borel measurable function such that $\int_\Omega f\,d\mu$ exists. Define $\nu(A) = \int_A f\,d\mu$, $A \in \mathcal{F}$. Then ν is countably additive; thus $(\Omega, \mathcal{F}, \nu)$ is a measure space.

Proof. Let A be a disjoint sequence of sets in $\mathcal{F}$. Then $\nu\left(\bigcup_{n=1}^{\infty} A_n\right) = \int_{\bigcup_{n=1}^{\infty} A_n} f\,d\mu = \sum_{n=1}^{\infty} \int_{A_n} f\,d\mu = \sum_{n=1}^{\infty} \nu(A_n)$, since μ is countably additive, as is $\int$.

Now for the generalization. First measurable, suppose $f = \sum_{i=1}^{n} a_i 1_{A_i}$ where the A_i are disjoint sets in $\mathcal{F}$. If f is almost and $0 < a_i < \infty$, then

$$\int_\Omega f\,d\mu = \sum_{i=1}^{n} a_i \int_{A_i} 1\,d\mu$$

by what we have proved, for nonnegative simple functions

$$= \sum_{i=1}^{n} a_i \int_{A_i} d\mu$$

by [illegible] which is the definition of the integral.

Take the [illegible] $\int_\Omega f\,d\mu = \sum_{i=1}^{n} a_i \mu(A_i)$.

Now $A_i \in \mathcal{F}$ since $f = a_i$ on A_i by [illegible] $\mu(A_i) \le \mu(\Omega)$. If $\mu(A_i) = \infty$ for some i, we are finished. So assume all $\mu(A_i)$ finite. Fix $\varepsilon > 0$. It follows from [illegible] and the fact that f is a sum of a finite number of simple functions, a simple [illegible] we can find a simple function g, $0 \le g \le f$, such that

$$\int_\Omega f\,d\mu \ge \int_\Omega g\,d\mu - \varepsilon, \qquad g = b_i \text{ on } B_i, \quad i = 1, \ldots, m.$$

Now

$$\mu(B_1 \cap \cdots \cap B_m) = \int_{B_1 \cap \cdots \cap B_m} d\mu \ge \int_{\bigcap_{i=1}^{m} A_i} d\mu = \sum_{i=1}^{n} \int_{A_i} d\mu$$

by induction we have proved for nonnegative simple functions. Hence

$$\lambda(B_1 \cup \cdots \cup B_n) = \sum_{i=1}^{n} \int_{B_i} 1\, d\mu \quad \text{and} \quad \lambda(E) = \sum_{i=1}^{\infty} \lambda(E_i) \quad \text{as claimed.}$$

Since $\lambda(E) = \lambda(E \cap B_i)$ and $E \cap B_i \subseteq B_i$, we have

$$\lambda(E) = \sum_{i=1}^{\infty} \nu(E \cap B_i).$$

Finally, let $h = h^+ - h^-$ be an arbitrary Borel measurable function. Then $\nu(E) = \int_E h^+ d\mu - \int_E h^- d\mu$ and $\int_E h^+ d\mu < \infty$ and $\int_E h^- d\mu < \infty$, so the result follows. ∎

The proof of 2.6.2 shows that ν is the difference of two measures λ^+ and λ^-, where $\lambda^+(B) = \int_B h^+ d\mu$ and $\lambda^- = \int_B h^- d\mu$. At least one of the measures λ^+ and λ^- must be finite.

2.6.3 Monotone Convergence Theorem. Let $h_1, h_2, \ldots$ form an increasing sequence of nonnegative Borel measurable functions, and let $h(x) = \lim_{n\to\infty} h_n(x)$ for all x. Then $\int_E h\, d\mu = \lim_{n\to\infty} \int_E h_n\, d\mu$. (Here each $\int_E h_n\, d\mu$ increases with n by 1.5.3(b); for short, $0 \le h_n \uparrow h$ implies $\int_E h_n\, d\mu \uparrow \int_E h\, d\mu$.)

Proof. By 1.5.3(b), $\int_E h_n\, d\mu \le \int_E h\, d\mu$ for all n, hence $\lambda = \lim_{n\to\infty} \int_E h_n\, d\mu \le \int_E h\, d\mu$. Let $0 < \delta < 1$, and set s be a nonnegative, finite-valued simple function with $s \le h$. Let $E_n = \{x : h_n(x) \ge \delta s(x)\}$. Then $E_n \subseteq E_{n+1}$ since $h_n \le h_{n+1}$, and $\bigcup_n E_n = E$ is σ-additive. Now $\int_E h_n\, d\mu \ge \int_{E_n} h_n\, d\mu$ by 1.5.3(b) and $\int_{E_n} h_n\, d\mu \ge \delta \int_{E_n} s\, d\mu$ by 1.5.3(a) and (b). By 1.6.1 and 1.2.5, $\int_{E_n} s\, d\mu \to \int_E s\, d\mu$, hence $\lambda \ge \delta \int_E s\, d\mu$. Thus $\lambda \ge \int_E s\, d\mu$ whenever $s \le h$ and s is simple. ∎

2.6.4 Additivity Theorem. Let f and g be Borel measurable, and assume that $f + g$ is well-defined. If $\int_E f\, d\mu$ and $\int_E g\, d\mu$ exist and $\int_E f\, d\mu + \int_E g\, d\mu$ is well-defined (not of the form $\infty - \infty$ or $-\infty + \infty$), then

$$\int_E (f + g)\, d\mu = \int_E f\, d\mu + \int_E g\, d\mu.$$

In particular, if f and g are integrable, so is $f + g$.

PROOF. If f and g are nonnegative simple functions, the s immediate from the definition of the integral. Assume f and g are nonnegative Borel measurable, and let $\{s_n\}$, $\{t_n\}$ be nonnegative simple functions increasing to f and g, respectively. Then $0 \le s_n + t_n \uparrow f + g$. Above $\int_E s_n\,d\mu = \int_E s_n\,d\mu + \int_E t_n\,d\mu$ by what we have proved for nonnegative simple functions; hence by Lem. 2, $\int_E (f+g)\,d\mu = \int_E f\,d\mu + \int_E g\,d\mu$.

Now if $f \ge 0$, $g \le 0$, $h = f - g \ge 0$ for x not in a fixed set N of measure 0, we have $f = h + (-g)$ where $\int_E f\,d\mu = \int_E h\,d\mu + \int_E -g\,d\mu$. If $\int_E g\,d\mu$ is finite, then $\int_E h\,d\mu = \int_E f\,d\mu + \int_E g\,d\mu$ and if $\int_E g\,d\mu = -\infty$, then since $h \ge 0$

$$\int_E f\,d\mu \ge \int_E h\,d\mu = \infty,$$

contradicting the hypothesis that $\int_E f\,d\mu + \int_E g\,d\mu$ is well defined. Similarly, if $f \ge 0$, $g \le 0$, $h = f - g \ge 0$, we obtain $\int_E h\,d\mu = \int_E f\,d\mu - \int_E g\,d\mu$ by replacing all functions by their negatives. (Explicitly, $-g \ge 0$, $-f \le 0$, $-h = -f - (-g)$, $-f \le 0$, and the above argument applies.)

Let

$$\begin{aligned}
E_1 &= \{x : f(x) > 0,\ g(x) < 0\}, \\
E_2 &= \{x : f(x) > 0,\ g(x) \le 0,\ h(x) > 0\}, \\
E_3 &= \{x : f(x) > 0,\ g(x) = 0,\ h(x) < \infty\}, \\
E_4 &= \{x : f(x) = 0,\ g(x) < 0,\ h(x) < \infty\}, \\
E_5 &= \{x : f(x) = 0,\ g(x) = 0,\ h(x) = 0\}.
\end{aligned}$$

The above argument shows that $\int_{E_1} h\,d\mu = \int_{E_1} f\,d\mu - \int_{E_1} g\,d\mu$. Now $\int_{E_2} f\,d\mu = \sum_{n=1}^{\infty} \int_{E_2} f\,d\mu$, $\int_{E_2} g\,d\mu = \sum_{n=1}^{\infty} \int_{E_2} g\,d\mu$, so that $\int_{E_2} f\,d\mu + \int_{E_2} g\,d\mu = \sum_{n=1}^{\infty} \int_{E_2} h\,d\mu$, and since each $\int_{E_2} h\,d\mu$ exists, $\int_{E_2} h\,d\mu$ exists and $\int_{E_2} h\,d\mu = \int_{E_2} f\,d\mu - \int_{E_2} g\,d\mu$. If $\int_{E_3} h\,d\mu$ exists, then $\int_{E_3} h\,d\mu$ and $\int_{E_3} g\,d\mu$ are real, so $\int_{E_3} f\,d\mu$ exists. Thus $\int_{E_3} h\,d\mu = \int_{E_3} f\,d\mu - \int_{E_3} g\,d\mu$.

If this is not the case, $\int_{E_3} h\,d\mu = \int_{E_3} f\,d\mu - \int_{E_3} g\,d\mu = \infty$ for we have $\int_{E_3} f\,d\mu < \infty$, so that $\int_{E_3} g\,d\mu = -\infty$. $\int_{E_4} h\,d\mu = -\infty$. But then $\int_{E_4} f\,d\mu = 0$, $\int_{E_4} g\,d\mu = -\infty$; hence $\int_{E_4} f\,d\mu$ or $\int_{E_4} g\,d\mu = \infty$. (Note that $\int_{E_4} f\,d\mu < \infty$, $\int_{E_4} g\,d\mu < \infty$.) Similarly $\int_{E_5} f\,d\mu = \int_{E_5} g\,d\mu = -\infty$, and this is a contradiction. ∎

3.6.4 *Corollary.* (a) If $\{f_n\}_{n=1}^{\infty}$ are nonnegative Borel measurable,

$$\int_X \left(\sum_{n=1}^{\infty} f_n\right) d\mu = \sum_{n=1}^{\infty} \int_X f_n \, d\mu.$$

Thus any series of nonnegative Borel measurable functions may be integrated term by term.

(b) If $f = \sum_{n=1}^{\infty} f_n$ is Borel measurable, f is integrable iff $\sum_{n=1}^{\infty} \int |f_n| \, d\mu < \infty$.

(c) If f and g are Borel measurable with $|f| \le |g|$ a.e. and g is integrable, then f is integrable.

Proof. (a) $\sum_{n=1}^{N} f_n \uparrow \sum_{n=1}^{\infty} f_n$, and the result follows from 3.6.3 and 3.6.2.

(b) Since $|f| \le \sum_{n=1}^{\infty} |f_n|$, this follows from the definition of the integral and (a).

(c) By 3.6.3(b), $|g|$ is integrable, and the result follows from (b) above. ∎

A condition is said to hold *almost everywhere* with respect to the measure μ (or μ-*almost everywhere*, or μ-*a.e.*) if there is a set $E \in \mathcal{M}$ of measure 0 such that the condition holds outside of E. From the point of view of integration theory, functions that differ only on a set of measure 0 are indistinguishable. This is established in the following result.

3.6.5 *Theorem.* Let f, g, and h be Borel measurable functions.

(a) If $f = 0$ a.e. $[\mu]$, then $\int_E f \, d\mu = 0$.

(b) If $f = g$ a.e. $[\mu]$ and $\int_E f \, d\mu$ exists, then so does $\int_E g \, d\mu$ and $\int_E g \, d\mu = \int_E f \, d\mu$.

Proof.

(a) If $f = \sum_{j=1}^{n} a_j \chi_{A_j}$ is simple, then $a_j \mu(A_j \cap E) = 0$ by hypothesis, hence $\int_E f \, d\mu = 0$. If $f \ge 0$ and $s \le f$ is simple, then $s = 0$ a.e. $[\mu]$, hence $\int_E s \, d\mu = 0$, thus $\int_E f \, d\mu = 0$. If $f = f^+ - f^-$, then f^+ and f^-, being less than or equal to $|f|$, are 0 a.e. $[\mu]$, and the result follows.

(b) Let $A = \{x : f(x) \ne g(x)\}$. Then $g = g\chi_{A^c} + g\chi_A = h\chi_{A^c} + g\chi_A = f\chi_{A^c} + g\chi_A$. Since $g\chi_A = 0$ except on A, a set of measure 0, the result follows from (a) and 3.6.3(a). ∎

There are many integration theorems we may freely use for almost everywhere functions. For example, if $\{f_n\}$ is a sequence of nonnegative Borel measurable functions converging a.e. to a Borel measurable function f, then $\int_E f_n \, d\mu \to \int_E f \, d\mu$.

Another example: If g and h are Borel measurable and $g \ge h$ a.e., then $\int_E g \, d\mu \ge \int_E h \, d\mu$ [in the sense of 3.6.5(b)].

2.6.6 Theorem. Let f be Borel measurable.

(a) If f is integrable, then f is finite a.e.

(b) If $f \geq 0$ and $\int_E f\,d\mu = 0$, then $f = 0$ a.e. on E.

Proof. (a) Let $A = \{x : f(x) = \infty\}$. If $\mu(A) > 0$, then $\int_A f\,d\mu \geq \int_A n\,d\mu = n\mu(A) = \infty$, a contradiction.

(b) Let $B = \{x : f(x) > 0\}$. $B = \bigcup_{n=1}^{\infty} B_n$, where $B_n = \{x : f(x) > 1/n\}$. We have $0 = \int_E f\,d\mu \geq \int_{B_n} f\,d\mu \geq \frac{1}{n}\mu(B_n)$ by (2.5.9)(a). $\int_{B_n} f\,d\mu = 0$. But $\int_{B_n} f\,d\mu \geq (1/n)\mu(B_n)$, so that $\mu(B_n) = 0$ for all n, and thus $\mu(B) = 0$. ∎

The monotone convergence theorem was proved under the hypothesis that all functions were nonnegative. This assumption can be phrased somewhat differently as follows:

2.6.7 Extended Monotone Convergence Theorem. Let $g_1, g_2, \ldots, g$ be Borel measurable.

(a) If $g_n \geq h$ for all n, where $\int_E h\,d\mu > -\infty$, and $g_n \uparrow g$, then

$$\int_E g_n\,d\mu \uparrow \int_E g\,d\mu.$$

(b) If $g_n \leq h$ for all n, where $\int_E h\,d\mu < \infty$, and $g_n \downarrow g$, then

$$\int_E g_n\,d\mu \downarrow \int_E g\,d\mu.$$

Proof. (a) If $\int_E g_n\,d\mu = -\infty$, then by (2.6.6)(b) $\int_E g_k\,d\mu = -\infty$ for some k, and $\int_E g\,d\mu = -\infty$. Thus assume $\int_E g_n\,d\mu > -\infty$ for all n. By (2.6.6)(a) h is a.e. finite; change it to 0 on the set where it is infinite. Then $h = g_n - (g_n - h) \leq g_n$ a.e., hence by (2.6.4) $\int_E h\,d\mu \leq \int_E g_n\,d\mu$. The result follows from (2.6.3). (We must check that the additivity theorem usually applies, since $\int_E g_n\,d\mu = \infty$, $\int_E g\,d\mu$ and $\int_E g_n\,d\mu$ exist and are finite, then $= \infty$ by (2.6.6). Also $\int_E h\,d\mu$ is finite, so that $\int_E g_n\,d\mu = \int_E h\,d\mu + \int_E (g_n - h)\,d\mu$ are well defined.)

(b) $-g_n \uparrow -g$, $\int_E -h\,d\mu < \infty$, and $-g_n \geq -g$. By part (a), $-\int_E g_n\,d\mu \uparrow -\int_E g\,d\mu$, so $\int_E g_n\,d\mu \downarrow \int_E g\,d\mu$. ∎

The extended monotone convergence theorem covers the main applications. The limit of the integrals of a sequence of functions is the integral of the limit function. More general theorems of this type can be obtained by

we replace limits by upper or lower limits. If $f_1, f_2, \ldots$ are functions from E to $\overline{\mathbb{R}}$, $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$ are defined pointwise, that is,

$$\left(\liminf_{n\to\infty} f_n\right)(x) = \sup_{n} \inf_{k \geq n} f_k(x),$$

$$\left(\limsup_{n\to\infty} f_n\right)(x) = \inf_{n} \sup_{k \geq n} f_k(x).$$

2.6.5 *Fatou's Lemma.* Let $f_1, f_2, \ldots$ be Borel measurable.

(a) If $f_n \geq f$ for all n, where $\int_E f\,dx > -\infty$, then

$$\liminf_{n\to\infty} \int_E f_n\,dx \geq \int_E \left(\liminf_{n\to\infty} f_n\right) dx.$$

(b) If $f_n \leq F$ for all n, where $\int_E F\,dx < \infty$, then

$$\limsup_{n\to\infty} \int_E f_n\,dx \leq \int_E \left(\limsup_{n\to\infty} f_n\right) dx.$$

Proof (a) Let $g_n = \inf_{k \geq n} f_k$, $g = \liminf f_n$. Then $g_n \geq f$ for all n, $\int_E f\,dx > -\infty$, and $g_n \uparrow g$. By 2.6.3, $\int_E g\,dx = \lim_{n\to\infty} \int_E g_n\,dx$. But $g_n \leq f_n$, so

$$\lim_{n\to\infty} \int_E g_n\,dx = \liminf_{n\to\infty} \int_E g_n\,dx \leq \liminf_{n\to\infty} \int_E f_n\,dx.$$

(b) We may write

$$\int_E \left(\limsup_{n\to\infty} f_n\right) dx = -\int_E \liminf_{n\to\infty} (-f_n)\,dx$$

$$\leq -\liminf_{n\to\infty} \int_E (-f_n)\,dx \quad \text{by (a)}$$

$$= \limsup_{n\to\infty} \int_E f_n\,dx. \quad \blacksquare$$

The following result is immediate from Fatou's lemma and is often the most convenient way to verify the hypothesis of the dominated convergence theorem.

2.6.6 *Dominated Convergence Theorem.* If $f_1, f_2, \ldots, f, g$ are Borel measurable, $|f_n| \leq g$ for all n, where g is nonnegative and $\int_E g\,dx < \infty$, and $f_n \to f$ a.e., then f is integrable and $\int_E f_n\,dx \to \int_E f\,dx$.

Proof. We have $f = g$ a.e. hence f is integrable by 1.6.4(c). By 1.6.8,

$$\int_E \liminf f_j \, d\mu \le \liminf \int_E f_j \, d\mu \le \limsup \int_E f_j \, d\mu$$
$$\le \int_E \limsup f_j \, d\mu.$$

By hypothesis, since $\lim f_j = \liminf f_j = \limsup f_j = f$ a.e., so all terms of the above inequalities are equal to $\int_E f \, d\mu$. ∎

1.6.10 *Corollary.* If $f, f_1, f_2, \ldots$ are Borel measurable, $|f_j| \le g$ for all j, where $\int g < \infty$ (a nonnegative g is $< \infty$ a.e.), and $f_j \to f$ a.e., then $|f| < \infty$ a.e., f is μ integrable and $\int_E |f_j - f| \, d\mu \to 0$ as $j \to \infty$.

Proof. We have $|f_j|^p \le |g|^p$ for all j, so $|f|^p \le |g|^p$, and therefore $|f|^p$ is integrable. Also $|f_j - f|^p \le (|f_j| + |f|)^p \le (2|g|)^p$, which is integrable, and the result follows from 1.6.9. ∎

We have seen in 1.6.5(c) that $g \ge 0$ implies $\int_E g \, d\mu \le \int_E h \, d\mu$ and in fact $\int_E g \, d\mu < \int_E h \, d\mu$ for all $A \in \mathcal{F}$. There is a converse to this result.

1.6.11 *Theorem.* If μ is σ-finite on $\mathcal{F}$, and g and h are Borel measurable, $\int_E g \, d\mu$ exists and $\int_E g \, d\mu \le \int_E h \, d\mu$ for all $A \in \mathcal{F}$, then $g \le h$ a.e. [μ].

Proof. It is sufficient to prove this when μ is finite. Let

$$A_n = \left\{ x : g(x) \ge h(x) + \frac{1}{n}, \quad |h(x)| \le n \right\}.$$

Then

$$\int_{A_n} g \, d\mu \ge \int_{A_n} g \, d\mu \ge \int_{A_n} h \, d\mu + \frac{1}{n} \mu(A_n).$$

But

$$\int_{A_n} h \, d\mu \ge -n\mu(A_n),$$

and thus we have $\mu(A_n) \le 2n\mu(A_n)$. Since μ is finite, $\mu(A_n) < \infty$, hence $\mu(A_n) = 0$. The set $\{x : g(x) > h(x)\}$ is the union of the sets A_n, $n = 1, 2, \ldots$, and $\mu(A_n) = 0$. Consequently $g \le h$ a.e. [μ]. Clearly $g = h$ a.e. [μ] if $|h(x)| < \infty$ a.e. ∎

$\{x : \limsup f_n(x) = \infty\}$ and $\{x : \liminf f_n(x) = -\infty\}$, $g(x) = -\infty$ if $\{x : \liminf f_n(x) = -\infty\}$, we obtain

$$\int_X g\,d\mu \le \int_X f\,d\mu \le \int_X g\,d\mu + \int_X h\,d\mu = \int_X g\,d\mu + \int_X h\,d\mu.$$

Hence $\int_X f\,d\mu = \int_X g\,d\mu$. Then $\int_X f\,d\mu = \int_X g\,d\mu$ as desired.

$$\int_X f\,d\mu = \int_X g\,d\mu + \int_X h\,d\mu = \int_X g\,d\mu + \int_X h\,d\mu = \int_X g\,d\mu.$$

The following is an important result.

If f and g are integrable, the proof is simple. Let $S = \{x : g(x) = h(x)\}$. Then $\int_S f\,d\mu = \int_S h\,d\mu = \int_S g\,d\mu$ because S has measure zero. Thus by Theorem 1.6.3, $\int_X f\,d\mu = \int_X g\,d\mu$. Now, if $f = g$ a.e. on X, then $f = g$ on $X \setminus S$, where S has measure zero. Hence $\int_X f\,d\mu = \int_X g\,d\mu$. In this case, f need not be finite.

The reader may have noticed that several elementary theorems in Riemann integration theory were proved by starting with nonnegative simple functions and working up to nonnegative measurable functions and hence to arbitrary measurable functions. This technique is quite basic and will be used frequently. A good illustration of the method is the following result, which minimizes the notion of a measure preserving transformation, a key concept in ergodic theory. In fact, it is convenient also to state the indicators before proceeding to nonnegative simple functions.

6.6.13 Theorem. Let $T : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a measurable mapping, and let μ be a measure on $\mathcal{F}$. Define a measure $\mu_T = \mu T^{-1}$ on $\mathcal{F}'$ by

$$\mu_T(A') = \mu(T^{-1}(A')) \qquad A' \in \mathcal{F}'.$$

If $\Omega_1 = \Omega$, $\mathcal{F}_1 = \mathcal{F}$, and $\mu_1 = \mu$, T is said to preserve the measure μ.

If $f : (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $A \in \mathcal{F}'$, then

$$\int_{T^{-1}(A)} f(T(x))\,d\mu(x) = \int_A f(x)\,d\mu_T(x)$$

in the sense that if one of the integrals exists, so does the other, and the two integrals are equal.

Proof. If f is an indicator I_B, the desired conclusion states that

$$\mu(T^{-1}(A \cap B)) = \mu_T(A \cap B),$$

which is true by definition of $\int_A f\,d\mu$. If f is a nonnegative simple function $\sum_{i=1}^{n} c_i I_{A_i}$, then

$$\int_{A\cup B} f\,d\mu = \sum_{i=1}^{n} c_i \int_{A\cup B} I_{A_i}\,d\mu \qquad \text{by 1.6.1}$$

$$= \sum_{i=1}^{n} c_i \int_{A} I_{A_i}\,d\mu + \sum_{i=1}^{n} c_i \int_{B} I_{A_i}\,d\mu$$

$$= \int_A f\,d\mu + \int_B f\,d\mu \qquad \text{by what we have proved for indicators}$$

$$= \int_{A\cup B} f\,d\mu \qquad \text{by 1.6.3.}$$

If f is nonnegative Borel measurable function, let $f_1, f_2, \ldots$ be nonnegative simple functions increasing to f. Then $\int_{A\cup B} f_n\,d\mu = \int_A f_n\,d\mu + \int_B f_n\,d\mu$ by what we have proved for simple functions, and the monotone convergence theorem gives the desired result for f.

Finally, if $f = f^+ - f^-$ is an arbitrary Borel measurable function, we have proved that the result holds for f^+ and f^-, say, $\int_{A\cup B} f^+\,d\mu = \int_A f^+\,d\mu + \int_B f^+\,d\mu$. If, say, $\int_{A\cup B} f^+\,d\mu < \infty$, then $\int_{A\cup B} f^-\,d\mu < \infty$, and it follows that it is valid to subtract the integrals term by term, and the resulting integrals sum to $\int_{A\cup B} f\,d\mu$.

There is a very difficult, yet very important, result concerning measurable functions of integration. It is often helpful to start with indicators and work upward. In fact it is possible to suspect that almost everything can be proved this way. This, of course, there are exceptions. For example, you will run into trouble trying to prove the proposition "All students are indicators."

We shall adopt the following convention: $(\Omega, \mathcal{A}, \mu)$ is a measure space and A is an interval $[a, b]$, $[a, \infty)$, $(-\infty, b]$, or $(-\infty, \infty)$. The integral $\int_A f(x)\,dx$ or $\int_a^b f(x)\,dx$, if it exists, will also be denoted by $\int_a^b f(x)\,dx$ and $\int_a^b f(x_1, \ldots, x_n)\,dx_1 \cdots dx_n$ when we are integrating on $\mathbb{R}^n$. The endpoints may be deleted from the interval without changing the integral, since the Lebesgue measure of a single point is 0. If f is integrable with respect to μ, then we say that f is *Lebesgue integrable*. A different notation, such as $\int_a^b f(x)\,F(dx)$, will be used for the Riemann–Stieltjes integral of f with respect to F.

Problems

The next seven problems give conditions under which some of the most commonly occurring operations in real analysis may be performed: taking a limit under the integral sign, integrating an infinite series term by term, and differentiating under the integral sign.

1. Let $f = f(x, y)$ be a real-valued function of two real variables defined for $a \le x \le b$, $c \le y \le d$. Assume that for each x, $f(x, y)$ is a Borel measurable function of y, and that there is a Borel measurable $g : [c, d] \to \mathbb{R}$ such that $|f(x, y)| \le g(y)$ for all x, y, and $\int_c^d g(y)\,dy < \infty$. If $\varphi(x) = \int_c^d f(x, y)\,dy$ and $\lim_{x \to x_0} f(x, y)$ exists for all $y \in [c, d]$, show that

$$\lim_{x \to x_0} \int_c^d f(x, y)\,dy = \int_c^d \left(\lim_{x \to x_0} f(x, y) \right) dy.$$

2. For $f_1, f_2, \ldots$ a Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$, if

$$\sum_{n=1}^{\infty} \int_{\Omega} |f_n|\,d\mu < \infty,$$

then show that $\sum_{n=1}^{\infty} f_n$ converges a.e. $[\mu]$ to a finite-valued function and $\int_{\Omega} \left(\sum_{n=1}^{\infty} f_n \right) d\mu = \sum_{n=1}^{\infty} \int_{\Omega} f_n\,d\mu$.

3. Let $f = f(x, y)$ be a real-valued function of two real variables. Suppose for each x, $y \mapsto f(x, y)$ is a Borel measurable function of y for each fixed x. Assume that for each x, $f(x, \cdot)$ is integrable over $[a, b]$ with respect to Lebesgue measure. Suppose that the partial derivative $f_1(x, y)$ of f with respect to x exists for all (x, y), and a positive Borel measurable function $h : [a, b] \to [0, \infty)$ exists such that $|f_1(x, y)| \le h(y)$ for all x, y where $\int_a^b h(y)\,dy < \infty$.

   Show that $d/dx \int_a^b f(x, y)\,dy$ exists for all $x \in (a, b)$, and equals $\int_a^b f_1(x, y)\,dy$. It must be verified that $y \mapsto f_1(x, y)$ is a Borel measurable for each x.

4. Let μ be a measure on $(\Omega, \mathcal{F})$ and $A_1, A_2, \ldots$ a sequence of sets in $\mathcal{F}$. For Fatou's lemma: to show that

$$\mu\left(\liminf_n A_n \right) \le \liminf_n \mu(A_n).$$

   It is enough to show that

$$\mu\left(\bigcup_{n=k}^{\infty} \bigcap_{m=n}^{\infty} A_m \right) \le \liminf_{n \to \infty} \mu(A_n).$$

   Then if μ is finite and $A = \liminf_n A_n$ then $\mu(A) \le \liminf_{n \to \infty} \mu(A_n)$. (For another proof of this see Section 1.2, Problem 10.)

5. Given an open plot of a sequence of Lebesgue integrable functions f_n converging everywhere to a Lebesgue integrable function f, such that

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} f_n(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx.$$

Thus the hypotheses of the dominated convergence theorem and Fatou's lemma cannot be dropped.

6 (a) Show that $\int_0^\infty e^{-x}\sin nx\,dx = \lim_{\beta\to\infty}\int_0^\beta e^{-x}\sin nx\,dx$ for $n=1,2,\ldots$.

(b) Show that $\int_0^\infty e^{-x}\cos nx\,dx = \lim_{\beta\to\infty}\int_0^\beta e^{-x}\cos nx\,dx$ for $n=1,2,\ldots$.

7 If $(\Omega,\mathcal{S},\mu)$ is the completion of $(\Omega,\mathcal{S}_0,\mu_0)$ and f is a finite measurable function on $(\Omega,\mathcal{S})$, show that there is a Borel measurable function g on $(\mathbb{R},\mathcal{B}_1)$ such that $f=g$ except on a subset of a set in $\mathcal{S}_0$ of measure 0. ($\mathbb{R}$ and $\mathcal{B}_1$ are defined in Problem 5.)

8 If f is a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$ and $a=c_0<c_1<\cdots<c_n=b$, show that

$$\int_{f^{-1}((a,b])} f(x)\,dx = \int_{f^{-1}((c_0,c_1])} f(x)\,dx + \cdots + \int_{f^{-1}((c_{n-1},c_n])} f(x)\,dx$$

in the sense that if one integral exists so does the other, and the two are equal. (Here m will denote Lebesgue measure.)

1.7 Comparison of Lebesgue and Riemann Integrals

In this section we show that integration with respect to Lebesgue measure is more general than Riemann integration, and we obtain a precise criterion for Riemann integrability.

Let $[a,b]$ be a bounded closed interval on the real line, and let f be a bounded real-valued function on $[a,b]$. As usual, we begin the discussion of the Riemann integral of f by considering partitions of $[a,b]$. A partition of $[a,b]$ is a finite set of points $P=\{x_0,x_1,\ldots,x_n\}$ such that $a=x_0\le x_1\le\cdots\le x_n=b$. We say sometimes that the points $x_0,x_1,\ldots,x_n$ themselves are the partition. We define the upper and lower sums of f relative to P as follows:

Let

$$M_i = \sup\{f(x) : x_{i-1}\le x\le x_i\}, \qquad i=1,\ldots,n,$$

$$m_i = \inf\{f(x) : x_{i-1}\le x\le x_i\}, \qquad i=1,\ldots,n,$$

and define step functions α and β called the upper and lower functions corresponding to P by

$$\alpha(x) = M_i \quad \text{if} \quad x_{i-1}\le x\le x_i, \qquad i=1,\ldots,n,$$

$$\beta(x) = m_i \quad \text{if} \quad x_{i-1}\le x\le x_i, \qquad i=1,\ldots,n.$$

$\alpha(x)$ and $\beta(x)$ may be chosen arbitrarily at $x_0,\ldots,x_n$. The upper and lower sums are then

$$U(P) = \sum_{i=1}^{n} M_i(x_i - x_{i-1}),$$

$$L(P) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}).$$

Now we take as a measure space $\Omega = [a, b]$, $\mathcal{A} = \mathcal{B}[a, b]$, the Borel sets on $[a, b]$, and $\mu = \lambda[a, b]$ = Lebesgue measure. Since α and β are simple functions, we have

$$U(P) = \int_a^b \alpha \, d\lambda, \qquad L(P) = \int_a^b \beta \, d\lambda.$$

Now let $P = P_1, P_2, \ldots$ be a sequence of partitions of $[a, b]$ such that P_{i+1} is a refinement of P_i for each i, and such that $\|P_i\|$ (the length of the largest subinterval of P_i) approaches zero as $i \to \infty$. If α_i and β_i are the upper and lower functions corresponding to P_i, then

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq f \geq \cdots \geq \beta_2 \geq \beta_1.$$

Thus α_i and β_i approach limit functions α and β. If f is bounded by M, then all α_i and β_i may be bounded by M as well, and the hypothesis that f is Riemann integrable on $[a, b]$ with respect to μ states

$$\mu[a, b] = 0 \quad \text{a.e.}$$

By the monotone convergence theorem,

$$\lim_{i \to \infty} U(P_i) = \lim_{i \to \infty} \int_a^b \alpha_i \, d\lambda = \int_a^b \alpha \, d\lambda,$$

or,

$$\lim_{i \to \infty} L(P_i) = \lim_{i \to \infty} \int_a^b \beta_i \, d\lambda = \int_a^b \beta \, d\lambda.$$

We shall need one other fact readily that if α is not an equivalent to any of the subintervals of P_i,

$$f \text{ is Riemann integrable on } [a, b] \quad \text{iff} \quad \alpha(x) = f(x) = \beta(x).$$

This follows by a standard ε–δ argument.

If $\lim_{i \to \infty} U(P_i) = \lim_{i \to \infty} L(P_i) = s$ for some number s, independent of the particular sequence of partitions P_i as long as $\|P_i\| \to 0$, then f is said to be Riemann integrable on $[a, b]$, and $s = s(f)$ is said to be the value of the Riemann integral of f on $[a, b]$. The above argument shows that f is Riemann integrable if

$$\int_a^b \alpha(x) \, dx = \int_a^b \beta \, d\mu = s.$$

independent of the particular sequence of partitions. If f is Riemann integrable, we write

$$\int_a^b f(x)\,dx = \int_a^b f\,dm.$$

We are now ready for the main results.

1.3.1 Theorem. Let f be a bounded real-valued function on $[a,b]$.

(a) The function f is Riemann integrable on $[a,b]$ iff f is continuous almost everywhere on $[a,b]$ (with respect to Lebesgue measure).

(b) If f is Riemann integrable on $[a,b]$, then f is integrable with respect to Lebesgue measure on $[a,b]$, and the two integrals are equal.

Proof. (a) If f is Riemann integrable,

$$\int_a^b f(x)\,dx = \int_a^b f\,dm = \int_a^b f\,du.$$

As $g \le f \le h$, $\int g\,du \le \int f\,du \le \int h\,du$; but $\int g\,du = \int g\,dm$ and $\int h\,du = \int h\,dm$. Since f is continuous a.e., $g(x) = f(x) = h(x)$ a.e.; hence f is measurable. Conversely, assume f is continuous a.e. Then $g = f = h$ a.e. Since g and h are limits of simple functions, and hence are Borel measurable, there is $E \in \mathcal{B}$ differing from a measurable partition by a subset of a set of measure 0, and therefore f is measurable because of the completeness of the measure space. (See Section 1.5, Problem 4.) Since f is bounded, it is integrable with respect to Lebesgue measure, $g = f = h$ a.e., and moreover

$$\int_a^b g\,du = \int_a^b f\,dm = \int_a^b h\,du. \tag{1}$$

independent of the particular sequence of partitions. Therefore, f is Riemann integrable.

(b) If f is Riemann integrable, then f is measurable and Eq. (1) holds, and $\int f\,du = \int f\,dm$, the desired result. ∎

Theorem 1.3.2. The results equally well in n dimensions with $[a,b]$ replaced by a closed bounded interval in $\mathbb{R}^n$. The proof is essentially the same.

A somewhat more complicated but informative way to topologize integrals over the interval of integration or infinite or discontinuous f is introduced. Some results are given in Problem 3.

We have seen the convergence conditions that allow the interchange of limit operations and Lebesgue integrals. Sections 1.6 (for example, see Problems 1–3 of Section 1.6) has corresponding results for Riemann integrals

functions are more complicated, essentially because the limit of a sequence of Riemann integrable functions need not be Riemann integrable, even if the limit sequence is monotonically convergent (see Problem 4). This Riemann integrability of the limit function must be assumed in hypotheses and this is a serious limitation on the scope of the results.

Problems

1 The function defined on $[0, 1]$ by $f(x) = \cdots$ if x is rational, and $f(x) = \cdots$ if x is irrational, is the standard example of a function that is Lebesgue integrable but is not Riemann integrable. But what is wrong with the following reasoning?

If we consider the behavior of f on the individual intervals of a partition, the constant value 1 and 0 is irrelevant, since we can ignore sets of measure zero. Since the rationals have Lebesgue measure zero, f is therefore continuous almost everywhere and hence is Riemann integrable.

2 Let f be a bounded real-valued function on the finite interval $[a, b]$. Let h be an increasing right-continuous function on $[a, b]$ with corresponding Lebesgue–Stieltjes measure μ defined on the Borel subsets of $[a, b]$.

Define M_i, m_i, w_i, and $\bar{f}$, $\underline{f}$, and ω by

$$M_i = \sup\{f(x) : x_{i-1} \le x \le x_i\}, \qquad m_i = \inf\{f(x) : x_{i-1} \le x \le x_i\},$$

$$w_i = M_i - m_i, \qquad \bar{f}(x) = \sum_{i=1}^{n} M_i \chi_{[x_{i-1}, x_i)}(x), \qquad \underline{f}(x) = \sum_{i=1}^{n} m_i \chi_{[x_{i-1}, x_i)}(x),$$

$$\omega(x) = \sum_{i=1}^{n} w_i \chi_{[x_{i-1}, x_i)}(x),$$

where χ_E indicates the characteristic function of the set E. Let $\{P_k\}$ be a sequence of partitions of $[a, b]$ such that P_{k+1} is a refinement of P_k, with $\bar{f}_k$ and $\underline{f}_k$ the upper and lower functions corresponding to P_k.

Show that

$$\lim_{k \to \infty} L(P_k) = \int_a^b \underline{f}\, d\mu,$$

$$\lim_{k \to \infty} U(P_k) = \int_a^b \bar{f}\, d\mu.$$

Where $\bar{f} = \lim_{k \to \infty} \bar{f}_k$, $\underline{f} = \lim_{k \to \infty} \underline{f}_k$. If $U(P_k)$ and $L(P_k)$ approach the same limit as $\|P_k\| \to 0$, independent of the particular sequence of partitions $\{P_k\}$, this number is called the *Riemann–Stieltjes integral* of f with respect to $F = \cdots$ on $[a, b]$, and f is said to be *Riemann–Stieltjes integrable* with respect to F on $[a, b]$.

(a) Show that f is Riemann–Stieltjes integrable if f is continuous on $[a, b]$.

(b) Show that if f is Riemann–Stieltjes integrable then f is integrable with respect to the completion of the measure μ, and the two integrals are equal.

3. If $f : [a, \infty) \to \mathbb{R}$ the improper Riemann integral of f may be defined as

$$\int_a^\infty f(x)\,dx = \lim_{b \to \infty} \int_a^b f(x)\,dx$$

if this limit exists and is finite.

(a) Show that if f has an improper Riemann integral, it is continuous a.e. [Lebesgue measure] on $[a, \infty)$, but not conversely.

(b) If f is nonnegative and has an improper Riemann integral, show that f is integrable with respect to the completion of Lebesgue measure and the two integrals are equal. Give a counterexample to this result if f is not nonnegative. Hypothesis is required.

4. Give an example of a sequence of functions f_n on $[a, b]$ such that each f_n is Riemann integrable, $f_n \le 1$ for all n, $f_n \to f$ pointwise, but f is not Riemann integrable.

Note. References on measure and integration will be given at the end of Chapter 2.

CHAPTER

2

FURTHER RESULTS IN MEASURE AND INTEGRATION THEORY

2.1 INTRODUCTION

This chapter consists of a variety of applications of the basic integration theory developed in Chapter 1. Perhaps the most important result is the Radon–Nikodym theorem, which is fundamental in modern probability theory and other parts of analysis. It will be instructive to consider a special case of this result before proceeding to the general theory. Suppose that F is a distribution function on $\mathbb{R}$, and assume that F has a jump of magnitude $p_k > 0$ at the point x_k, $k = 1, 2, \ldots$. That is, $p_k = F(x_k) - F(x_k-)$, where $F(x_k-) = \lim_{x \uparrow x_k} F(x)$. Let μ be the probability measure concentrated on $\{x_1, x_2, \ldots\}$ with $\mu\{x_k\} = p_k$ for all k, and let F be a distribution function corresponding to μ. Then $F(x) = \sum_{x_k \le x} p_k$ is a continuous step function except at the x_k's. The corresponding Lebesgue–Stieltjes measure λ_F satisfies $\lambda_F\{x_k\} = p_k$ for all k. Now, if only "countably many" values are taken on by F, then $F(x) = \sum_{k=1}^{\infty} p_k f(x_k)$, where f is the characteristic function of some countable set $\{x_1, x_2, \ldots\}$. It follows that $\lambda_F(A) = \sum_{x_k \in A} p_k f(x_k) = 0$ for all Borel sets A. Thus the characteristic function of $\{x : F'(x) = 0\}$ is 1 a.e. $[\lambda_F]$. That is, $F'(x) = 0$ a.e. $[\lambda_F]$. Conversely, if $F'(x) = 0$ a.e. $[\lambda_F]$, then $\lambda_F\{x : F'(x) = 0\} = \lambda_F(\mathbb{R}) = 1$. Thus $F'(x) = 0$ a.e. $[\lambda_F]$ if and only if λ_F is the Lebesgue–Stieltjes measure determined by 0, in other words, $F = 1_{\mathbb{R}}$.

The natural conjecture is that if F is a measure on $\mathcal{B}(\mathbb{R})$ and $F'(x) = 0$ for all x, then we can write $F(B) = \int_B p(x)\,dx$. But $p(x) = 0$, for some nonnegative Borel measurable p. However, as Theorem 1.5.3 will show, the conjecture is false unless the hypothesis is strengthened. Not only must we assume that F is a finite measure on $\mathcal{B}(\mathbb{R})$, but we must assume that F is absolutely continuous with respect to Lebesgue measure μ, that is, if $\mu(B) = 0$, then $F(B) = 0$. In general, F may be represented as the sum of two measures λ_1 and λ_2, where λ_1 is absolutely continuous with respect to μ and λ_2 is singular with respect to μ, which means that λ_2 is concentrated on a set of Lebesgue measure 0. A simple example of a measure singular with respect to μ is

one that is concentrated on a countable set; however, as we shall see, more complicated examples exist.

The first step in the development of the general Radon–Nikodým theorem is the Jordan–Hahn decomposition, which represents a countably additive real function as the difference of two measures.

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. A Borel measurable function φ such that $\int_\Omega \varphi \, d\mu$ exists, $\lambda(A) = \int_A \varphi \, d\mu$ defines, for $A \in \mathcal{A}$, a countably additive set function on $\mathcal{A}$. We call λ the *indefinite integral* of φ (with respect to μ). If μ is Lebesgue measure and $\varphi = [\text{a continuous function}]$, then $\lambda(A) = \int_A \varphi(x)\,dx$ is the familiar improper integral of calculus. A substantial part of the proof of 5.1.2 is the existence of two measures, at least one of which is finite, that are singular to those that any countably additive set function can be represented in this way. First, a preliminary result.

5.1.1 Theorem. *Let λ be a countably additive extended real valued set function on the σ-field $\mathcal{A}$ of subsets of Ω. That is, assume that λ never takes on both $+\infty$ and $-\infty$ as values, that $\lambda(\emptyset) = 0$, and that $\lambda(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \lambda(A_n)$ for any sequence $\{A_n\}$ of disjoint sets in $\mathcal{A}$.*

Then either $\lambda(B) \geq 0$ for all $B \in \mathcal{A}$ or $\lambda(B) \leq 0$ for all $B \in \mathcal{A}$.

Before giving the proof, let us look at some special cases. If λ is a measure, the result is trivial; take $P = \Omega$, $N = \emptyset$. If λ is the indefinite integral of φ with respect to μ, we argue that

$$\lambda(A) = \int_A \varphi \, d\mu = \int_{A \cap \{\varphi \geq 0\}} \varphi \, d\mu + \int_{A \cap \{\varphi < 0\}} \varphi \, d\mu.$$

Thus

$$\int_{A \cap \{\varphi < 0\}} |\varphi| \, d\mu \leq \lambda(A) \leq \int_{A \cap \{\varphi \geq 0\}} |\varphi| \, d\mu.$$

Therefore we may take $P = \{x : \varphi(x) \geq 0\}$, $N = \{x : \varphi(x) < 0\}$.

Proof of 5.1.1. We may assume that λ never takes on $+\infty$ (otherwise we work with $-\lambda$). Let A_1 be a set with $\lambda(A_1) = m > 0$. If $A_1 = \bigcup_{n=1}^\infty A_{1n}$ where each A_{1n} is at most $m/2$, we may remove all but one disjoint subset A_{1n_1}, where each A_{1n} is at most $\lambda(A_{1n_1}) \geq m/2$. Continuing, we obtain $A_{11} \supset A_{12} \supset \cdots \supset A_{1k}$ with $\lambda(A_{1k}) \geq m/2^{k-1}$ and $\lambda(A_1 - A_{1k}) \geq 0$. For example, if $n = 3$, we have (with intersections written as products)

$$A_1 = A_{11}A_{12}A_{13} \cup A_{11}A_{12}A_{13}' \cup A_{11}A_{12}'A_{13} \cup A_{11}A_{12}'A_{13}' \cup A_{11}'A_{12}A_{13} \cup A_{11}'A_{12}A_{13}' \cup A_{11}'A_{12}'A_{13} \cup A_{11}'A_{12}'A_{13}'$$

where A_{1j}' denotes $A_1 - A_{1j}$.

Let $B_1 = \bigcup_{n=1}^\infty A_{1n}$. If $\lambda(B_1) = 0$ or $\lambda(A_1 - B_1) = 0$ for all n, we are finished. If not, there is a subset of some A_{1n}, say $A_2 = A_{1n}$, with $\lambda(A_2) \geq m/2$ and $\lambda(A_1 - A_2) \geq 0$.

definition of A_n. Also, if $n' > n$, each $A_{n,k}$ is either a subset of a set in $\mathcal{A}_{n'}$ or disjoint from it (for example, $A_{n,k}\cap A_{n',k'} \neq \emptyset$ and $A_{n,k}\cap A_{n',k'}$ is the union of $A_{n,k}\cap A_{n',k'}$ and $A_{n,k}\setminus A_{n',k'}$). Thus $\bigcup_{n=1}^{\infty}\mathcal{A}_n$ can be expressed as a union of $\mathcal{B}_1$ and sets in $\mathcal{A}_1$ disjoint from $\mathcal{B}_1$, so $\mu(\bigcup_{n=1}^{\infty}\mathcal{A}_n) \geq \lambda$. (Note that, for example, if $\mathcal{A}_1 = \{A_{1,1}, A_{1,2}\}$ and $\mathcal{A}_2 = \{A_{2,1}, A_{2,2}\}$, it may happen that $A_{2,1}\cap A_{1,1} = \emptyset$, but $A_{2,1}\cap A_{1,2}$ is not, so the argument that $\mu(\mathcal{A}_1) \geq \lambda$ is not necessarily correct.) Consequently,

$$\lambda = \mu(\mathcal{A}_1) \leq \mu(\mathcal{A}_n) \leq \lambda\left(\bigcup_{k=1}^{n} A_{n,k}\right) \to \lambda\left(\bigcup_{k=1}^{\infty} A_{n,k}\right) \quad \text{as} \quad n \to \infty \quad \text{by (1.7)(b)}.$$

Let $\mathcal{B}_n = \bigcup_{k=1}^{n} A_{n,k}$. Now $\bigcup_{n=1}^{\infty}\mathcal{B}_n \in \mathcal{A}$, and for all n, $\lambda(\bigcup_{n=1}^{\infty}\mathcal{B}_n) = \infty$ for all n by (1.7)(b), $\lambda(\bigcup_{n=1}^{\infty}\mathcal{B}_n) = \lambda(\mathcal{B}_n) \to \lambda(\mathcal{A}_1)$ as $n \to \infty$. Thus

$$\sup\{\lambda = \mu(A) : A \in \mathcal{A}\} \leq \lim_{n\to\infty} \lambda\left(\bigcup_{k=1}^{n} A_{n,k}\right) = \lambda(\mathcal{A}_1) \leq \sup\{\lambda = \mu(A) : A \in \mathcal{A}\},$$

hence $\sup\{\lambda = \mu(A) : A \in \mathcal{A}\} = \lambda(\mathcal{A}_1)$. The above argument applies to any particular $\mathcal{A}_1 \in \mathcal{F}$ with $\lambda(\mathcal{A}_1) = \inf\{\lambda : \mathcal{A} \in \mathcal{F}\}$. ∎

We now prove the main theorem of this section.

1.8.7 *Jordan–Hahn Decomposition Theorem.* Let μ be a countably additive extended real valued set function on the σ-field $\mathcal{A}$ of sets in Ω. Let

$$\lambda^+(\mathcal{A}) = \sup\{\mu(B) : B \in \mathcal{A}, B \subset \mathcal{A}\},$$
$$\lambda^-(\mathcal{A}) = -\inf\{\mu(B) : B \in \mathcal{A}, B \subset \mathcal{A}\}.$$

Then λ^+ and λ^- are measures on $\mathcal{A}$ such that $\mu = \lambda^+ - \lambda^-$.

Proof. We may assume μ never takes on the value $+\infty$. (For if $\mu = +\infty$ on $\mathcal{A}$ for some $\mathcal{A} \in \mathcal{A}$, then $\lambda^-(\mathcal{A}) = \infty$ by definition of countably additive set functions. Thus $-\mu$ never takes on the value $+\infty$. But $\lambda^+ = (-\mu)^-$ and $(-\mu)^- = \mu^+$, so that if the theorem is proved for $-\mu$ it holds for μ as well.

Let $\mathcal{B}$ be a set in $\mathcal{A}$ which is μ-null and μ-antinull, as in 2.1.1. Since $\mu(\emptyset) = 0$, we have $\lambda^+(\mathcal{B}) = \lambda^-(\mathcal{B}) = 0$. We claim that

$$\lambda^+(\mathcal{A}\cap\mathcal{B}) \leq 0, \qquad \lambda^-(\mathcal{A}\cap\mathcal{B}^c) \leq 0 \qquad \text{for all} \qquad \mathcal{A} \in \mathcal{A}. \tag{*}$$

For if $\lambda^+(\mathcal{A}\cap\mathcal{B}) > 0$, then $\lambda^+(\mathcal{B}) = \lambda^+(\mathcal{A}\cap\mathcal{B}) + \lambda^+(\mathcal{A}^c\cap\mathcal{B})$. Since $\lambda^+(\mathcal{B})$ is finite, so are $\lambda^+(\mathcal{A}\cap\mathcal{B})$ and $\lambda^+(\mathcal{A}^c\cap\mathcal{B})$. Hence $\lambda^+(\mathcal{A}^c\cap\mathcal{B}) = \lambda^+(\mathcal{B}) - \lambda^+(\mathcal{A}\cap\mathcal{B})$

$\nu(B)$, contradicting the fact that $\lambda(B) = \infty$. If $\lambda(A) < \infty$, then $\lambda(B) \geq (\lambda - \varepsilon)(A) = \nu(B) + \nu(A \setminus B) - \varepsilon(B)$, a contradiction.

We now prove that

$$\lambda^{+}(A) = \sup\{\lambda(B) : B \subseteq A\} \qquad \nu^{+}(A) = -\inf\{\nu(B) : B \subseteq A\} \tag{3}$$

The theorem will follow from (3). We have, for $B \subseteq A$,

$$\begin{aligned} \nu(B) &= \lambda(B) - \varepsilon(B) \\ &\geq \lambda(B) - \varepsilon(A) \\ &= \lambda(B) + \lambda(A \setminus B) - \lambda(A) \\ &= \lambda(A) - \lambda(A \setminus B) \end{aligned}$$

Thus $\lambda^{+}(A) = \lambda(A) - \inf\{\lambda(A \setminus B) : B \subseteq A\}$. By definition of λ^{+} proving the first assertion. Similarly,

$$\begin{aligned} \nu(B) &= \lambda(B) - \varepsilon(B) \\ &= \nu(B) + \varepsilon(B) \\ &\geq \nu(B) + \varepsilon(A \setminus B) \\ &= \nu(A) - \nu(A \setminus B) \end{aligned}$$

Thus $-\nu^{+}(A) = \nu(A) - \inf\{\nu(A \setminus B) : B \subseteq A\}$. By definition of ν^{+} completing the proof. ∎

2.2.3 *Corollaries.* Let λ be a countably additive extended real-valued set function on the σ-field $\mathcal{F}$.

(a) The set function ν is the difference of two measures, at least one of which is finite.

(b) If $\lambda(A)$ is finite $\lambda(B)$ is never ∞ or $-\infty$, then λ is bounded.

(c) There is a set $B \in \mathcal{F}$ such that $\lambda(A \cap B) \geq 0$ and $\lambda(A \cap B^{c}) \leq 0$ for all $A \in \mathcal{F}$.

(d) If there exists $\alpha > 0$ such that $|\lambda(B)| < \infty$ and $|\lambda(A \cap B)| \geq \alpha$ for all $A \subseteq B$, then $|\lambda(A)| = \infty$, where $A \subseteq B$ means $A \in \mathcal{F}$ and $A \subseteq B$.

(e) If $\mathcal{F}$ is a σ-field of subsets of Ω such that $\lambda(A) < \infty$ for all $A \in \mathcal{F}$, then $\lambda(A) < \infty$, where $\Omega = A \cup A^{c}$.

Proof. (a) If λ is finite, then $\lambda = \lambda^{+} - \lambda^{-}$ is finite. If $\lambda = \infty$, λ^{+} is finite [see (b).]

(b) In 2.1.2, λ^{+} and λ^{-} are both finite valued for any $A \in \mathcal{F}$, since $\lambda^{+}(A) \leq \lambda^{+}(\Omega) + \lambda^{-}(\Omega) < \infty$.

(a) ... as follows from (a) of 2.1.2 and ...

(b) Repeat the proof of 2.1.2 after Eq. (3).

(c) By (b), $\lambda(A) = \lambda(A \cap E^+) + \lambda(A \cap E^-)$, $A \in \mathcal{F}$. Take $A = E^+ \cap E^-$ to obtain $\lambda^+(E^+ \cap E^-) = 0$. Assume $\lambda^+(A) = \lambda(A \cap E^+) > 0$ for some $A \in \mathcal{F}$. Since $A = E^+ \cap A$ we obtain $\lambda^+(E^+ \cap A) > 0$. Therefore $\lambda^+(A \cap A) > 0$. The corresponding statement for $\lambda^-(A) = -\lambda(A \cap E^-)$ shows that $\lambda^-(A \cap A) = 0$. Use (a) to conclude the following.

Corollary 2.1.3 (Jordan decomposition) *Every signed measure λ is the difference of two mutually singular measures λ^+ and λ^-.*

2.1.4 *Terminology.* We call λ^+ the upper variation or positive part of λ, λ^- the lower variation or negative part of λ, and $|\lambda| = \lambda^+ + \lambda^-$ the total variation of λ. Since $\lambda = \lambda_1 - \lambda_2$ implies that $|\lambda(A)| \le |\lambda_1|(A) + |\lambda_2|(A)$, $A \in \mathcal{F}$, we have a useful result; see Problem 4.

Note that $\lambda^+ = \lambda^-$ if and only if $\lambda(A) = 0$ for all $A = E \cap B$, $B \in \mathcal{F}$.

The preceding signed measure is sometimes used for the difference of two measures. By 2.1.3 (a), this is equivalent to a real vector space with a vector lattice structure.

Problems

1. Let μ be an arbitrary probability measure on $\mathcal{B}(\mathbb{R}^1)$ and let Q be point mass at 0. That is, $Q(B) = 1$ if $0 \in B$, $Q(B) = 0$ if $0 \notin B$. Find the Jordan–Hahn decomposition of the signed measure $\lambda = P - Q$.

2. Let $\nu(A) = \int_A f \, d\mu$, $A \in \mathcal{F}$, in the σ-finite measure space $(\Omega, \mathcal{F}, \mu)$, where $\int_\Omega f \, d\mu$ exists. Thus ν is a signed measure on $\mathcal{F}$. Show that

$$\nu^+(A) = \int_A f^+ \, d\mu, \qquad \nu^-(A) = \int_A f^- \, d\mu, \qquad |\nu|(A) = \int_A |f| \, d\mu.$$

3. If a signed measure λ on the σ-field $\mathcal{F}$ is the difference of two measures λ_1 and λ_2, show that $\lambda^+ \le \lambda_1$, $\lambda^- \le \lambda_2$.

4. Let λ be a signed measure on the σ-field $\mathcal{F}$. Show that $|\lambda(A)| = \sup\{\sum_{n=1}^\infty |\lambda(A_n)| : A_1, A_2, \ldots$ are disjoint measurable subsets of A, $\bigcup_{n=1}^\infty A_n = A\}$. If λ_1 and λ_2 are signed measures on $\mathcal{F}$, then $|\lambda_1 + \lambda_2| \le |\lambda_1| + |\lambda_2|$.

2.2 Radon–Nikodym Theorem and Related Results

If $(\Omega, \mathcal{F}, \mu)$ is a measure space, then $\lambda(A) = \int_A g \, d\mu$, $A \in \mathcal{F}$, defines a signed measure if $\int_\Omega g \, d\mu$ exists. If, moreover, $A \in \mathcal{F}$ and $\mu(A) = 0$, then $\lambda(A) = 0$. For $g \ge 0$ a.e. $[\mu]$, or, more generally, $g > -\infty$ a.e. $[\mu]$, and the result continues to hold

If μ is a measure on the σ-field $\mathcal{F}$ and λ is a signed measure on $\mathcal{F}$ we say that λ is *absolutely continuous with respect to* μ, or *dominated by* μ, if $\mu(A) = 0$ implies $\lambda(A) = 0$ for all $A \in \mathcal{F}$. Thus if λ is an indefinite integral with respect to μ, as in (5.2.6), then $\lambda \ll \mu$. The Radon–Nikodym theorem is a converse in the sense that if $\lambda \ll \mu$ and μ is σ-finite, then λ is an indefinite integral with respect to μ. As we shall see, the importance of this result is not limited to this theorem.

5.3.7 *Radon–Nikodym Theorem.* Let μ be a σ-finite measure and λ a signed measure on a σ-field $\mathcal{F}$ of subsets of Ω. Assume that λ is absolutely continuous with respect to μ. Then there is a Borel measurable function $g : \Omega \to \mathbb{R}$ such that

$$\lambda(A) = \int_A g\,d\mu \qquad \text{for all} \qquad A \in \mathcal{F}.$$

If h is another such function, then $g = h$ a.e. $[\mu]$.

Proof. The uniqueness statement follows from (5.2.1). We break the existence proof into several parts.

(a) Assume λ and μ are finite measures.

Let $\mathcal{H}$ be the set of all non-negative μ-integrable functions f such that $\int_A f\,d\mu \le \lambda(A)$ for all $A \in \mathcal{F}$. Formally order $\mathcal{H}$ by saying $f \le g$ a.e. $[\mu]$ for $f, g \in \mathcal{H}$. Let $\alpha = \sup\{\int f\,d\mu : f \in \mathcal{H}\}$. Then $\mathcal{H}$ is nonempty (since it contains the zero function) and α is finite (since μ and λ are finite measures). Choose $f_1, f_2, \ldots \in \mathcal{H}$ such that $\int f_n\,d\mu \to \alpha$. Formally order $\mathcal{H}$ by saying $f \le g$ a.e. $[\mu]$ for $f, g \in \mathcal{H}$. Let $\alpha = \sup_{f \in \mathcal{H}} \int f\,d\mu$. Then $\alpha < \infty$ since μ and λ are finite measures. Choose $f_1, f_2, \ldots \in \mathcal{H}$ such that $\int f_n\,d\mu \to \alpha$. Let $g_n = \max(f_1, \ldots, f_n)$. Then $g_n \in \mathcal{H}$, g_n increases to a limit g, and $\int g_n\,d\mu \to \alpha$ (use the monotone convergence theorem and the fact that $g_n \ge f_n$). We claim that $g \in \mathcal{H}$, and $\int g\,d\mu = \alpha$ (it will follow that g is a maximal element of $\mathcal{H}$). To show that g belongs to $\mathcal{H}$, let A be any set in $\mathcal{F}$. Then

$$\int_A g_n\,d\mu \le \lambda(A) \qquad \text{so} \qquad \int_A g\,d\mu = \lim_n \int_A g_n\,d\mu \le \lambda(A).$$

But $\int_A g_n\,d\mu \le \lambda(A)$ for all n, and therefore $\int_A g\,d\mu \le \lambda(A)$.

Now can you have our nonlinear element $f_1 \in \mathcal{F}^1$ $(A) = \nu(A) = \int_A g(x)\,dx$, $g \in \mathcal{F}$. Then λ_1 is a measure $\lambda_1 : \mathcal{A} \to [0,\infty]$ and $\lambda_1(\Omega) < \infty$. If λ is not identically 0, then $\lambda_1(\Omega) > 0$. Since

$$\mu(\Omega) - \nu(\Omega) > 0 \qquad \text{for some} \qquad n \geq 1 \tag{1}$$

Apply 2.1.5(a): ν is a signed measure on $\mathcal{A}$. So ν is a set $D \in \mathcal{F}$ such that for all $A \in \mathcal{F}$,

$$\mu(A \cap D) - \nu(A \cap D) \geq 0 \tag{2}$$

and

$$\mu(A \cap D^c) - \nu(A \cap D^c) \leq 0. \tag{3}$$

We claim that $\mu(D) > 0$. For if $\mu(D) = 0$, then $\lambda(D) = 0$ by absolute continuity. And therefore $\lambda(D) = 0$ by definition of λ. Take $A = \Omega$ in (3) to obtain

$$\begin{aligned} 0 &\geq \mu(D^c) - \lambda(D^c) \\ &= \mu(\Omega) - \mu(D) - \lambda(\Omega) + \lambda(D) \qquad \text{since} \qquad \mu(D) = \lambda(D) = 0 \\ &> 0 \qquad \text{by} \qquad (1). \end{aligned}$$

a contradiction. Hence $\mu(D) > 0$. Define $\mu(A) = \mu(A \cap D)$ and $\nu(A) = \nu(A \cap D)$.

$$\int_A f\,d\mu = \int_A g\,d\mu \quad \text{and} \quad \int_A f\,d\nu \leq \int_A g\,d\nu \qquad \forall A \in \mathcal{A}$$

$$\Rightarrow \int_A f\,d(\mu - \nu) \leq \int_A g\,d(\mu - \nu)$$

Thus $\int_A (f - g)\,d\mu \leq 0$. But $f - g > 0$ on the set D, with $\mu(D) > 0$, contradicting the minimality of g. This is the contradiction that we needed. It follows.

(b) Assume μ is a Lebesgue measure. ν not Lebesgue measure.

Let Ω be the disjoint union of sets A_n with $\lambda(A_n) < \infty$, and let $\lambda_n(A) = \lambda(A \cap A_n)$, $A \in \mathcal{A}$. By part (a) we have a nonnegative Borel measurable g_n with $\lambda_n(A) = \int_A g_n\,d\mu$, $A \in \mathcal{A}$. Then $\lambda(A) = \int_A g\,d\mu$, where $g = \sum_{n=1}^{\infty} g_n$.

(c) Assume μ is a finite measure. λ an arbitrary measure.

Let $\mathcal{F}$ be the class of sets $C = \mu$-a.e. such that $\lambda(C) < \infty$. Since λ is assumed to be σ-finite, $\mathcal{F} \neq \emptyset$. Note that $\mathcal{F} \subset \mathcal{A}$. So $\mathcal{F}$ is not empty. Let $\alpha = \sup\{\mu(C) : C \in \mathcal{F}\}$ and pick $C_n \in \mathcal{F}$ such that $\mu(C_n) \to \alpha$. If $C = \bigcup_{n=1}^{\infty} C_n$, then $C \in \mathcal{F}$ by definition of $\mathcal{F}$ and $\alpha \geq \mu(C) \geq \mu(C_n) \to \alpha$. Hence $\mu(C) = \alpha$

By part (b), there is a nonnegative $g: \Omega \to \mathbb{R}$, measurable relative to $\mathcal{F}$, and $\mathcal{A}(\mathcal{A})$ such that

$$\nu(A \cap C) = \int_{f^{-1}(\cdot)} g \, d\mu \qquad \text{for all} \qquad A \in \mathcal{H}$$

Now remember an arbitrary set $A \in \mathcal{H}$.

Case 1. Let $\mu(A \cap C) > 0$. Then $\nu(A \cap C) = \infty$, hence $g(\omega) \geq \alpha$ on, then $\mathbb{E}_\mu[g \mid \mathcal{F}] \geq \alpha$ since

$$\alpha \geq g(\omega) \geq \mathbb{E}_\mu[g \mid \mathcal{F}](\omega) = g(\omega) \quad \mu(A \cap C)\text{-a.e.} = \alpha(\omega) = \alpha,$$

a contradiction.

Case 2. Let $\mu(A \cap C) = 0$. Then $\nu(A \cap C) = 0$ by absolute continuity. Thus in either case $\int_{f^{-1}(\cdot)} g \, d\mu = 0$, it follows that

$$\nu(A) = \nu(A \cap C) = \nu(A \cap C^c) = \int_{C^c} g \, d\mu,$$

where $g = \infty$ on C, $g = \alpha$ on C^c.

(c) Assume μ is σ-finite measure, ν an arbitrary measure.

Let Ω be the union of disjoint sets A_n with $\mu(A_n) < \infty$. By part (a), there is a nonnegative function $g_n: A_n \to \mathbb{R}$, measurable with respect to $\mathcal{F}_n$ and $\mathcal{H}(\mathcal{F})$, such that $\nu(A \cap A_n) = \int_{A \cap A_n} g_n \, d\mu$, $A \in \mathcal{H}$. We may write this as $\nu(A \cap A_n) = \int_A g_n \, d\mu$, where $g_n(\omega)$ is taken as zero for $\omega \notin A_n$. Then $\nu(A) = \sum_n \nu(A \cap A_n) = \sum_n \int_{A \cap A_n} g_n \, d\mu = \int_A g \, d\mu$, where $g = \sum_n g_n$.

(d) Assume ν is a σ-finite measure and an arbitrary signed measure.

Write $\nu = \nu^+ - \nu^-$ as above, say ν^- is finite. By part (c) there are nonnegative Borel measurable functions g_1 and g_2 such that

$$\nu^+(A) = \int_A g_1(\omega) \, d\mu, \qquad \nu^-(A) = \int_A g_2(\omega) \, d\mu, \qquad A \in \mathcal{H}$$

Since ν^- is finite, g_2 is integrable hence by (a) $g_2 < \infty$ and $\nu(A) = \int_A (g_1 - g_2) \, d\mu$. ∎

2.2.2 *Corollaries.* Under the hypothesis of 2.2.1,

(a) If ν is finite then g is μ integrable, hence $g < \infty$ a.e. $[\mu]$.

(b) If $\mathcal{H}$ is finite so that Ω can be expressed as a countable union of sets A_n such that $\mathcal{H} \cap A_n$ is finite (equivalently $\mathcal{H}(A_n)$ is finite), then $g < \infty$ a.e. $[\mu]$.

(c) If $\mathcal{H}$ is countable, then $g < \infty$ a.e. $[\mu]$.

choice. A useful map is obtained by assuming the case of 2.2.1. More naturally, we may proceed as follows:

(a) Observe that $\lambda(G) = \int_X g\,d\mu$ is finite by hypothesis.

(b) By 2.1.5, g is finite a.e. $[\mu]$ on each A_n, hence finite a.e. $[\mu]$ on X.

(c) Let $A = \{x : g(x) = \infty\}$. Then $\lambda(A) = \int_A g\,d\mu \le \lambda(X) < \infty$. Thus $g\chi_A$ is a nonnegative function whose integral is 0, so that $g\chi_A = 0$ a.e. $[\mu]$ by 1.6.4(b). Since $g(x) > 0$ on A, we must have $\mu(A) = 0$. ∎

If $\nu(E) = \int_E g\,d\mu$ for each $E \in \mathcal{S}$, g is called the **Radon–Nikodym derivative** of ν with respect to μ, written $d\nu/d\mu$. If μ is Lebesgue measure, then g is often called simply the **derivative** of λ.

There are converse statements to 2.2.2(a) and (c). Suppose that

$$\lambda(E) = \int_E h\,d\mu, \qquad E \in \mathcal{S},$$

where $\int_E h\,d\mu$ is assumed to exist. If h is μ-integrable, then λ is finite. If g is finite $[\mu]$, then $\mu(A) = 0$, so that ν is μ-continuous. Note that ν-finiteness of μ is not assumed. However, the converse of 2.2.2(b) is false if μ is not ν-finite a.e. $[\nu]$, $[\lambda]$, and λ is not ν-finite (see Problem 1).

We now establish a property that is in a sense opposite to absolute continuity.

2.2.3 Definitions. Let μ and ν be measures on the σ-field $\mathcal{S}$. We say that μ is **singular** with respect to ν (written $\mu \perp \nu$) iff there is a set $A \in \mathcal{S}$ such that $\mu(A) = 0$ and $\nu(A^c) = 0$. In other words, μ is singular with respect to ν iff μ is singular with respect to ν relative to some measurable subspace. If λ_1 and λ_2 are signed measures on $\mathcal{S}$, we say that λ_1 and λ_2 are **mutually singular** iff $|\lambda_1| \perp |\lambda_2|$.

If $\mu_1 \perp \mu_2$, then $\mu_1(A) = \mu_2(A^c) = 0$ for some A, so that μ_1 essentially assigns positive measure to sets in $\mathcal{S}$. This is not true of absolutely continuous measures. On the other hand, if $\mu_1 \ll \mu_2$, μ_2 can have no effect on sets of μ_2-measure 0.

If λ is a signed measure with respective part λ^+ and negative part λ^-, we have $\lambda^+ \perp \lambda^-$ by 2.1.3(b) and (d).

Before establishing some results about absolute continuity and singularity, we need the following lemma. Although the proof is quite simple, the result is applied very often in analytic, especially probabilistic, arguments.

2.2.4 Borel–Cantelli Lemma. If $\lambda(E_n) < \infty$ for $n = 1, 2, \ldots$ and $\sum_{n=1}^{\infty} \lambda(E_n) < \infty$, then $\lambda(\limsup_n E_n) = 0$.

Lemma Recall that $\mathcal{E}$ is a semiring of subsets of Ω and μ is a measure on $\mathcal{E}$. Then

$$\mu(\limsup_{n\to\infty} A_n) \le \mu\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\right) \qquad \text{for all } m$$

$$\le \sum_{n=m}^{\infty} \mu(A_n) \to 0 \qquad \text{as } m \to \infty,$$

since $\sum_{n=1}^{\infty} \mu(A_n) < \infty$.

2.2.5 **Lemma** Let μ be a measure and λ_1 and λ_2 signed measures on the σ-field $\mathcal{F}$.

(a) If $\nu = \lambda_1 + \lambda_2$ and $\nu \ll \mu$, then $\lambda_1 \ll \mu$ and $\lambda_2 \ll \mu$.

(b) If $\lambda_1 \ll \mu$ and $\lambda_2 \ll \mu$, then $|\lambda_1 + \lambda_2| \ll \mu$ and $|\lambda_1| \ll \mu$.

(c) If $\lambda_1 \ll \mu$ and $\lambda_1 \perp \mu$, then $\lambda_1 = 0$.

(d) If $\lambda_1 \ll \mu$ and $\lambda_2 \perp \mu$, then $\lambda_1 \perp \lambda_2$.

(e) If λ_1 is finite, then $\lambda_1 \ll \mu$ iff $\lim_{\mu(A)\to 0} \lambda_1(A) = 0$.

Proof. (a) Let $\mu(A) = \mu(B) = 0$, $A \in \mathcal{F}^{\lambda_1}$, $B \in \mathcal{F}^{\lambda_2}$, $A \cap B = \emptyset$. Then $\mu(A \cup B) = 0$ and $\lambda_1(C) = \lambda_2(C) = 0$ for every $C \in \mathcal{F}^{\nu}$ with $C \subset A^c \cap B^c$; hence $\nu = \lambda_1 + \lambda_2 = 0$.

(b) Let $\mu(A) = 0$. If $\sum |\lambda_1(A_j)| > 0$, then (see 2.1.2) $\lambda_1(B) > 0$ for some $B \subset A$, since $\lambda_1(\emptyset) = 0$. This is a contradiction. Hence $|\lambda_1| \ll \mu$, and similarly $|\lambda_2| \ll \mu$, so $|\lambda_1 + \lambda_2| \ll \mu$ by (a). Since $\lambda \ll \mu$ iff $\lambda^+ \ll \mu$ and $\lambda^- \ll \mu$, this may also be proved using Section 2.1, Problem 4(g). The converse is clear.

(c) Let $\mu(A) = 0$, $|\lambda_1|(B) = 0$. By (b), $|\lambda_1|(A) = 0$ so $|\lambda_1| = 0$.

(d) By $\mu \perp \lambda$, there is a set $A \in \mathcal{F}^{\lambda_1}$ with $\lambda_1(A) = \lambda_1(A^c) = 0$. Thus $\lambda_1 = 0$.

(e) If $\mu(A_n) < 2^{-n}$ implies $\lambda_1(A_n) > \varepsilon$, and $\mu(A) = 0$ set $A_n = A$ in case all of the A_n are distinct. Then $\lambda_1(A) = 0$, so $\lambda \ll \mu$.

Conversely let $\lambda_1 \ll \mu$.

If $\lim_{\mu(A)\to 0} \lambda_1(A) \neq 0$ we can find $\varepsilon > 0$ and sets $A_n \in \mathcal{F}^{\lambda_1}$ with $\mu(A_n) < 2^{-n}$ and $|\lambda_1(A_n)| > \varepsilon$ for all n. Let $A = \limsup_n A_n$. By 2.2.4, $\mu(A) = 0$. But $\lambda_1(\bigcup_{n=m}^{\infty} A_n) \ge \lambda_1(A_m) \ge \varepsilon$ for all m, hence by (a) $\lambda_1(A) \ge \varepsilon$, a contradiction to (c). Thus $\lim_{\mu(A)\to 0} \lambda_1(A) = 0$, and the result follows since $\lambda_1(A) = \lambda_1(A \cap B)$.

If λ is not indefinite integral with respect to μ (hence $\lambda \not\ll \mu$), then 2.2.5(e) gives an easy proof. If $\lambda(B) = \int_B f\,d\mu$, $A = \{x : f(x) > 0\}$, then

$$\int_A f\,d\mu = \int_{A\cap\{f>0\}} f\,d\mu + \int_{A\cap\{f\le 0\}} f\,d\mu \le \lambda(A) < \int_A f\,d\mu,$$

By 1.6 and 1.23(b), $\int_{E_n \cup F_n} |f|\,d\mu$ can be made less than $\varepsilon/2$ for large n, say $n \geq N$. Since $\nu = \lambda$ and $\mu(E_n) < \varepsilon/2n$, we see that $\int_{E_n} f\,d\mu < \varepsilon/2$ also.

If μ is a measure and λ is a signed measure on the σ-field $\mathcal{M}$, λ may or may not be absolutely continuous or singular with respect to μ. However, as 2.14 will show, the two concepts of absolute continuity and singularity are enough to describe the relation between λ and μ. In fact, the set function λ can be written as the sum of two signed measures, one absolutely continuous and the other singular with respect to μ.

2.13 *Lebesgue Decomposition Theorem.* Let μ be a measure on the σ-field $\mathcal{M}$, λ a definite signed measure that is σ-finite with respect to μ. Then λ has a unique decomposition $\lambda = \lambda_1 + \lambda_2$, where λ_1 and λ_2 are signed measures such that $\lambda_1 \ll \mu$ and $\lambda_2 \perp \mu$.

Proof. First assume that λ is a finite measure and let $\mathcal{H} = \{h : \int_E h\,d\mu \leq \lambda(E)$ for every $E \in \mathcal{M}\}$. Since $0 \in \mathcal{H}$, $\mathcal{H}$ is not empty. If $h_1, h_2 \in \mathcal{H}$ and $A = \{x : h_1(x) \geq h_2(x)\}$, then $\int_E \max(h_1, h_2)\,d\mu = \int_{E \cap A} h_1\,d\mu + \int_{E \cap A^c} h_2\,d\mu \leq \lambda(E \cap A) + \lambda(E \cap A^c) = \lambda(E)$, so $\max(h_1, h_2) \in \mathcal{H}$. We show that $\sup_{h \in \mathcal{H}} \int h\,d\mu = \int h_0\,d\mu$ for some $h_0 \in \mathcal{H}$. For if $\int h_n\,d\mu \to \sup_{h \in \mathcal{H}} \int h\,d\mu$ and $h_0 = \max(h_1, \ldots, h_n)$, then

$$\int_E h_0\,d\mu \leq \int_E h_1\,d\mu + \cdots + \int_E h_n\,d\mu \leq \lambda(E),$$

a contradiction. Now define

$$\lambda_1(E) = \int_E h_0\,d\mu, \qquad \lambda_2(E) = \lambda(E) - \int_E h_0\,d\mu, \qquad E \in \mathcal{M}.$$

If $\mu(E) = 0$, then $\int_E h_0\,d\mu = 0$, so $\lambda_1(E) = 0$. Hence $\lambda_1 \ll \mu$. Since $\lambda_2(A) \geq 0$ and $\lambda_2(X) = \lambda(X) - \int h_0\,d\mu \geq 0$, we have $h_0 \leq \lambda$. Since $\lambda_2(A) \geq 0$ and $\lambda_2(E) \leq \lambda(E)$ for every $E \in \mathcal{M}$, we have $h_0 + \chi_A \in \mathcal{H}$ and $\lambda_2 \perp \mu$.

Now if λ is a definite signed measure, the above argument applies to λ^+ and λ^-, proving the existence of the desired decomposition.

To prove uniqueness, first assume λ finite. If $\lambda = \lambda_1 + \lambda_2 = \lambda_1' + \lambda_2'$ where $\lambda_1, \lambda_1' \ll \mu$ and $\lambda_2, \lambda_2' \perp \mu$, then $\alpha = \lambda_1 - \lambda_1' = \lambda_2' - \lambda_2$ is both absolutely continuous and singular with respect to μ; hence α is identically 0 by 2.12(c). If λ is σ-finite and $\{A_n\}$ is the disjoint union of sets A_n with $\lambda(A_n) < \infty$, apply the above argument to each μ_{A_n} and sum the results together to obtain uniqueness of λ_1 and λ_2. ∎

[illegible] W. Rudin, *Real and Complex Analysis*, New York, 1970.

Note that as a corollary of the proof, if λ is a finite measure on a measurable space $(\Omega, \mathcal{F})$, then λ is a finite signed measure, but $\lambda(\Omega)$ may be negative. Also, if λ is a probability measure, then $\lambda(\Omega) = 1$.

Problems

1. Give an example of a measure μ and a nonnegative, nonintegrable Borel measurable function g such that the measure λ defined by $\lambda(A) = \int_A g \, d\mu$ is not a finite measure.

2. If $\lambda(A) = \int_A g \, d\mu$, $A \in \mathcal{F}$, and g is an integrable, we know that λ is finite on $\mathcal{F}$. However, if g is only assumed to be finite a.e. μ, show that λ need not be σ-finite on $\mathcal{F}$. Show, however, that λ is σ-finite on the σ-field $\mathcal{F}_0 = \{A \in \mathcal{F} : \mu(A) = 0 \text{ or } \mu(A^c) = 0\}$. Construct an example to show that $\mathcal{F}_0$ may be trivial.

3. Give an example to show that the conclusion of the Radon–Nikodým theorem fails if either μ is not σ-finite but there is an $\mathcal{F}$-measurable g such that $\nu(A) = \int_A g \, d\mu$ for all $A \in \mathcal{F}$, or if ν is not σ-finite.

4. (Change of variable) Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and g a nonnegative Borel measurable function on Ω. Define a measure λ on $\mathcal{F}$ by

$$\lambda(A) = \int_A g \, d\mu, \qquad A \in \mathcal{F}.$$

   Show that if f is a Borel measurable function on Ω,

$$\int_A f \, d\lambda = \int_A f g \, d\mu$$

   in the sense that if one of the integrals exists, so does the other, and the two integrals are equal. (Intuitively, $d\lambda = g \, d\mu$, or $d\lambda = g(x) \, dx$.)

5. Show that Theorem 2.2 Step 1 and 2 are false if λ is not finite.

6. (Complex measures) If $(\Omega, \mathcal{F})$ is a measurable space, a complex measure ν on $\mathcal{F}$ is a countably additive complex-valued function on $\mathcal{F}$; that is, $\nu = \lambda_1 + i\lambda_2$, where λ_1 and λ_2 are finite signed measures.

   (a) Define the total variation of ν as

$$|\nu|(\Omega) = \sup\left\{\sum_{i=1}^{n} |\nu(E_i)| : E_1, \ldots, E_n \right.$$
$$\left. \text{disjoint measurable, } \bigcup_{i=1}^{n} E_i = \Omega, n = 1, 2, \ldots\right\}.$$

   Show that $|\nu|$ is a measure on $\mathcal{F}$. (This definition is consistent with the earlier notion of total variation of a signed measure; see Section 1.3, Problem 4.)

In the discussion below, λ is a real-valued function, μ is a positive measure (real signed measures or complex measures), and μ is the corresponding total variation measure. We define $\lambda \ll \mu$ (" λ is absolutely continuous with respect to μ") if $\mu(E) = 0$ implies $\lambda(E) = 0$. Similarly, $\lambda \perp \mu$ ("λ and μ are mutually singular") if $|\lambda| \perp |\mu|$. Establish the following results.

(a) If $|\lambda_1 + \lambda_2| \le |\lambda_1| + |\lambda_2|$, then $|\lambda_1 + \lambda_2| = |\lambda_1| + |\lambda_2|$ if and only if $|\lambda_1|(E)|\lambda_2|(E) \ge 0$ for every measurable set E. In particular, if $\lambda = \lambda_1 + i\lambda_2$ is a complex measure, then

$$|\lambda| = |\lambda_1| + |\lambda_2| \quad \text{if and only if} \quad |\lambda_1|(E)|\lambda_2|(E) = 0 \text{ for all } E \text{ by 2.1.3(a)}.$$

(b) If $\lambda \ll \mu$ and $\lambda \perp \mu$, then $\lambda = 0$.

(c) If $\lambda \ll \mu$, then $|\lambda| \ll \mu$, and conversely.

(d) If $\lambda \ll \mu$ and $\mu \ll \nu$, then $\lambda \ll \nu$.

(e) If $\lambda \ll \mu$ and $\lambda \perp \mu$, then $\lambda = 0$.

(f) If λ is finite, then $\lambda \ll \mu$ if and only if $\lim_{\mu(E)\to 0} \lambda(E) = 0$.

2.4 Absolutely Continuous Real Functions

We are going to apply the concepts of the previous section to study particular functions of a real variable. If $[a, b]$ is a closed bounded interval in $\mathbb{R}$ and $f : [a, b] \to \mathbb{R}$, f is said to be absolutely continuous if for each $\varepsilon > 0$ there is a $\delta > 0$ such that for all positive integers n and all families $(a_1, b_1), \ldots, (a_n, b_n)$ of disjoint open subintervals of (a, b) of total length $< \delta$, we have

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \varepsilon.$$

It is immediate that this property holds also for countably infinite families of disjoint open intervals of total length at most δ. It is also obvious from the definition that if f is continuous.

We can connect absolute continuity of functions with the earlier notion of absolute continuity of measures as follows.

2.4.1 Theorem. *Suppose that F and G are distribution functions on $[a, b]$ with corresponding positive Lebesgue-Stieltjes measures μ_F and μ_G. Let $f = F - G$, $\mu = \mu_F - \mu_G$, and let ν be the positive signed measure on $\mathcal{B}([a, b])$ with $\nu([c, d]) = f(d) - f(c)$. Let m be Lebesgue measure on $\mathcal{B}([a, b])$. Then $\mu \ll m$ if and only if F is absolutely continuous.*

Proof. Assume $\mu \ll m$. We claim that if $\varepsilon > 0$, then $\delta = \varepsilon/2$ will do. To see this, suppose $m(E) < \delta$. Applying 2.2.3(b) to μ and ν, we have $|\nu(E)| \le \mu(E)$. Thus if $(a_1, b_1), \ldots, (a_n, b_n)$ are disjoint

open intervals of total length at most δ,

$$\sum_{i=1}^{n} |f(b_i) - f(a_i)| = \sum_{i=1}^{n} |\mu(I_i)| \leq \sum_{i=1}^{n} \nu([a_i, b_i]) < \varepsilon.$$

Hence $\nu(\{0\}) = 0$ since $\mu(\{0\}) = 0$. Therefore f is absolutely continuous.

Now assume f absolutely continuous; $\varepsilon > 0$ choose $\delta > 0$ as in the definition of absolute continuity. If $\nu(\{0\}) = 0$, we must show that $\mu(\{0\}) = 0$. We use (4.1.1):

$$\nu(\{0\}) = \inf\{\mu(V) : V \supset \{0\}, \quad V \text{ open}\}$$

$$\mu_j(V) = \inf\{\alpha_j(I) : I \subset V, \quad I \text{ open}\}, \quad j = 1, 2.$$

(This problem assumes that the measures are defined on $\mathcal{B}(\mathbb{R})$ rather than $\mathcal{B}([a, b])$. The easiest way to do is to extend measures to $\mathcal{B}(\mathbb{R})$ by assigning measure 0 to $\mathbb{R} - [a, b]$.) Since μ is continuous, for each n we can find a decreasing sequence $\{V_n\}$ of open sets such that $\mu(V_n) > \mu(\{0\})$ and $\nu(V_n) < \nu(\{0\}) = 0$.

Choose a sequence $\{I_{n,j}\}$ so that $\bigcup_j I_{n,j} \supset V_n$ if V_n is the disjoint union of the open intervals $\{I_{n,j} : j = 1, 2, \ldots\}$. Then $\mu(V_n) \leq \sum_j |\mu(I_{n,j})|$. But f is continuous, hence

$$\mu(\{0\}) = \lim_{n \to \infty} \mu(V_n) \leq \lim_{n \to \infty} \sum_j |f(I_{n,j})| = \lim_{n \to \infty} |f(V_n)| = 0.$$

Therefore

$$|\mu(V_n)| \leq \sum_j |\mu(I_{n,j})| = \sum_j |f(I_{n,j})| < \varepsilon.$$

Since ε is arbitrary and $\mu(V_n) \to \mu(\{0\})$, we have $\mu(\{0\}) = 0$. ∎

If $f : [a, b] \to \mathbb{R}$ absolutely continuous and f as defined exactly as above. If λ and $\mathcal{G}$ are bounded distribution functions on $\mathbb{R}$ with corresponding Lebesgue–Stieltjes measures μ_1 and μ_2, then $f = F - G = (\lambda - \mu_1) - (\mu_2 - \mu_1)$ is the signed measure on $\mathcal{B}([a, b])$ that f is absolutely continuous if μ_1 is absolutely continuous with respect to Lebesgue measure, the proof is the same as in 2.3.7.

Any absolutely continuous function on $[a, b]$ can be represented as the difference of two absolutely continuous increasing functions. We prove this in the exercises at the end.

If $f:[a,b]\to\mathbb{R}$ and $P=\{x_0,x_1,\dots,x_n\}$ is a partition of $[a,b]$ define

$$V(f;P)=\sum_{i=1}^{n}|f(x_i)-f(x_{i-1})|.$$

The sup of $V(f;P)$ over all partitions of $[a,b]$ is called the variation of f on $[a,b]$. Various notations: $V_f(a,b)$, $V_a^b(f)$, $\bigvee_a^b f$, $T_a^b(f)$, or simply $V(a,b)$ if f is understood. We say that f is of bounded variation on $[a,b]$ if $V(f;[a,b])<\infty$. If $a<c<b$ a small argument shows that $V_a^b f=V_a^c f+V_c^b f$.

3.3.2 *Lemma.* If $f:[a,b]\to\mathbb{R}$ and f is absolutely continuous on $[a,b]$, then f is of bounded variation on $[a,b]$.

Proof. Choose $\varepsilon=1$ and let $\delta>0$ be chosen as in the definition of absolute continuity. If P is any partition of $[a,b]$, there is a refinement Q of P consisting of subintervals of length less than $\delta/2$. If $P=\{x_0,\dots,x_n\}$ and $Q=\{y_0,\dots,y_m\}$ with $m>n$, let k be the largest integer such that $x_{k-1}=y_{j-1}$ for some j, and let j be the largest integer such that $y_{j-1}=x_{k-1}$. Note that $y_{j-1}-x_{k-1}\le\delta/2$. Also $y_j-x_{k-1}<\delta$. Hence by construction of Q, $y_j\le x_k$. Thus

$$|f(y_j)-f(x_{k-1})|\le 1\quad\text{and}\quad |f(x_k)-f(y_j)|\le 1+\sum_{i=j+1}^{m}|f(y_i)-f(y_{i-1})|.$$

By absolute continuity, $V(f;Q)\le M+1$. But $V(f;P)\le V(f;Q)$ since the refining set can never decrease the variation. The result follows.

It is remarked that a monotone function f on $[a,b]$ is of bounded variation, $V_a^b(f)=|f(b)-f(a)|$. Thus if $f=F-G$, where F and G are increasing, then f is of bounded variation. The converse is also true.

3.3.3 *Lemma.* If $f:[a,b]\to\mathbb{R}$ and f is of bounded variation on $[a,b]$, then there are increasing functions F and G on $[a,b]$ such that $f=F-G$. If f is absolutely continuous, F and G may also be taken to be absolutely continuous.

Proof. Let $g(x)=V_a^x(f)$, $a\le x\le b$. If $a\le x<y\le b$, $V_a^y(f)=V_a^x(f)+V_x^y(f)\ge V_a^x(f)-|f(y)-f(x)|$. Thus $g(x)\ge f(x)$ and $g(y)\ge f(y)$. If $f(x)\ge f(y)$, then g is also increasing, but if $x<y$, then

$$\begin{aligned}G(y)-G(x)&=g(y)-f(y)-g(x)+f(x)\\&=V_a^y(f)-V_a^x(f)-f(y)+f(x)\\&=V_x^y(f)-f(y)+f(x)\\&\ge 0\quad\text{by definition of }V_x^y(f).\end{aligned}$$

Now assume f is absolutely continuous. Let $\varepsilon > 0$, choose $\delta > 0$ as in the definition of absolute continuity, and suppose $(a_1, b_1), \ldots, (a_N, b_N)$ are disjoint open intervals with total length at most δ. If γ is a subdivision of $[a, b]$ and $\gamma = 1, 2, \ldots, N$, then

$$\sum_{\gamma} V(f; \gamma) < \varepsilon \qquad \text{by short-term continuity of } f,$$

where the sum is extended over all γ in which $\gamma_1, \ldots, \gamma_N$ is chosen. Thus

$$\sum_{\gamma} V(f; [a_\gamma, b_\gamma]) \le \varepsilon$$

in other words

$$\sum_{\gamma=1}^{N} |f(b_\gamma) - f(a_\gamma)| < \varepsilon.$$

Therefore f is absolutely continuous. Since sums and differences of absolutely continuous functions are absolutely continuous, G is also absolutely continuous. $\square$

We have seen that there is a close connection between absolute continuity and integrals via the Radon–Nikodym theorem. The converse of course does not hold, as follows.

2.3.4 Theorem. *Let $f : [a, b] \to \mathbb{R}$. Then f is absolutely continuous on $[a, b]$ if f is an indefinite integral, that is, iff*

$$f(x) - f(a) = \int_a^x g(t)\,dt \qquad \text{for } a \le x \le b,$$

where $g : [a, b] \to \mathbb{R}$ is Borel measurable and integrable with respect to Lebesgue measure.

Proof. First assume f absolutely continuous. By 2.3.3, it is sufficient to assume f increasing. If μ is the Lebesgue–Stieltjes measure corresponding to f, and m is Lebesgue measure, then $\mu \ll m$ by 2.3.1. By the Radon–Nikodym theorem, there is an integrable function g such that $\mu(A) = \int_A g\,dm$ for all Borel subsets A of $[a, b]$. Take $A = [a, x]$ to obtain $f(x) - f(a) = \int_a^x g(t)\,dt$.

Conversely, assume $f(x) - f(y) = \int_y^x g(t)\,dt$. It is sufficient to assume g is real-valued, since we can treat g_1 and g_2 separately. Define $\nu([a,b]) = f(b) - f(a)$. As in (a), ν extends to a measure on $\mathbb{R}$, and its distribution function corresponding to μ_g is absolutely continuous by 2.10. But

$$F(x) - F(y) = \nu((y,x]) = \int_y^x g(t)\,dt = f(x) - f(y) \quad \text{for } y < x.$$

Therefore f is absolutely continuous on $\mathbb{R}$.

If g is Lebesgue integrable on $\mathbb{R}$, but not on $[a,\infty)$, part of the proof of 2.14 shows that the function defined by $\int_a^x g(t)\,dt$ is absolutely continuous, hence uniformly continuous, on $\mathbb{R}$. Another way of proving uniform continuity is to observe that

$$\int_{-\infty}^{x+h} g(t)\,dt - \int_{-\infty}^{x} g(t)\,dt = \int_{-\infty}^{x+h} g(t)\chi_{(x,x+h)}(t)\,dt$$

$\to 0$ as $h \to 0$, and this approaches 0 as $h \to 0$ by the dominated convergence theorem.

If $f(x) - f(y) = \int_y^x g(t)\,dt$, $a \le y \le x \le b$, and g is continuous at x, then f is differentiable at x and $f'(x) = g(x)$, by the proof given in calculus courses. In the present hypothesis is weakened, we can prove that $f'(x) = g(x)$ for almost every $x \in [a,b]$. One approach to this result is via the theory of differentiation of measures, which we now describe.

2.15 *Definition.* For the remainder of this section, μ is a signed measure on the Borel sets of $\mathbb{R}^k$, assumed finite on compact sets, that is, $\mu(K) < \infty$ if K is compact. If μ is Lebesgue measure, we denote, for each $x \in \mathbb{R}^k$,

$$(D\mu)(x) = \limsup_{r \to 0} \frac{\mu(C_r)}{m(C_r)} \qquad (\underline{D}\mu)(x) = \liminf_{r \to 0} \frac{\mu(C_r)}{m(C_r)}$$

where the C_r range over all open cubes of diameter less than r that contain x. It will be convenient (although not essential) to assume that the C_r have edges parallel to the coordinate axes.

We say that μ is *differentiable* at $x \in \mathbb{R}^k$ if $D\mu$ and $\underline{D}\mu$ are equal and finite at x, and we write $(D\mu)(x)$ for the common value. Then μ is differentiable at x if and only if for every sequence $\{C_r\}$ of open cubes containing x such that the diameter of C_r approaches 0, $\mu(C_r)/m(C_r)$ approaches a finite limit independent of the particular sequence.

The following result will play an important role.

2.3.6 *Vitali's theorem.* If $\{C_1, \ldots, C_N\}$ is a family of open balls in $\mathbb{R}^n$, there is a disjoint subfamily $\{C_{j_1}, \ldots, C_{j_k}\}$ such that $\sum_{i=1}^{k} \mathcal{L}^n(C_{j_i}) \geq 3^{-n} \sum_{i=1}^{N} \mathcal{L}^n(C_i)$.

Proof. Assume that the diameter of C_i decreases with i. Set $j_1 = 1$, and take j_2 to be the smallest index greater than 1 such that C_{j_2} is disjoint from C_{j_1}. If no such index exists, stop. Otherwise, take j_3 to be the smallest index greater than j_2 such that C_{j_3} is disjoint from $C_{j_1} \cup C_{j_2}$. Continue in this fashion to obtain disjoint sets $C_{j_1}, \ldots, C_{j_k}$. Now for any $i = 1, \ldots, N$, we have $C_i \cap C_{j_l} \neq \emptyset$ for some $j_l \geq j$, for if not, j_{l+1} would be the smallest index greater than j_l such that $C_{j_{l+1}}$ is disjoint from $C_{j_1} \cup \cdots \cup C_{j_l}$. Let C_i^* be the open ball with the same center as C_i and three times the radius, then $C_i \subset C_{j_l}^*$ and $\mathcal{L}^n(C_i^*) = 3^n \mathcal{L}^n(C_i)$. Therefore $C_i \subset C_{j_l}^*$ because C_i is a ball with the same center as $C_{j_l}^*$ and smaller radius. Thus $\bigcup_{i=1}^{N} C_i \subset \bigcup_{l=1}^{k} C_{j_l}^*$, and $\mathcal{L}^n\left(\bigcup_{i=1}^{N} C_i\right) \leq \sum_{l=1}^{k} \mathcal{L}^n(C_{j_l}^*) = 3^n \sum_{l=1}^{k} \mathcal{L}^n(C_{j_l})$. $\square$

We now prove the first differentiation result.

2.3.7 *Theorem.* Let μ be a Borel measure on $\mathbb{R}^n$ and λ a Radon measure on the Borel sets of $\mathbb{R}^n$. If $\mu(A) = 0$, then $D\mu = 0$ a.e. $[\lambda]$ on A.

Proof. Let $\varepsilon > 0$, let $E = \{x \in A : \overline{D}\mu(x) > \varepsilon\}$, and $\delta > 0$. Note that for each $x \in E$, $\inf\{\mu(C) : C \text{ closed}, x \in C\} = 0$, and it follows that E is a Borel set. Find an open set U containing E with $\mu(U) < \delta$. For each $x \in E$, there is an open ball C_x of diameter less than r with $x \in C_x \subset U$ and $\mu(C_x) > \varepsilon \lambda(C_x)$. By the Vitali covering theorem, E is covered by finitely many of these balls, say $C_1, \ldots, C_N$. If $\{C_{j_1}, \ldots, C_{j_k}\}$ is the subcollection of $\{C_i\}$ as in 2.3.6, we have

$$\lambda(E) \leq \lambda\left(\bigcup_{i=1}^{N} C_i\right) \leq 3^n \sum_{i=1}^{k} \lambda(C_{j_i}) \leq \frac{3^n}{\varepsilon} \sum_{i=1}^{k} \mu(C_{j_i})$$

$$\leq \frac{3^n}{\varepsilon} \mu\left(\bigcup_{i=1}^{k} C_{j_i}\right) \leq \frac{3^n}{\varepsilon} \mu(U) < \frac{3^n}{\varepsilon} \delta.$$

Where $E = \{x \in A : \overline{D}\mu(x) > \varepsilon\}$. Since ε is arbitrary, we have $\lambda(E) \leq 3^n \varepsilon^{-1} \delta$ for all $\delta > 0$. Thus $\lambda(E) = 0$. Take the sup over ε to obtain by 1.4.1, $\lambda(\{x \in A : \overline{D}\mu(x) > 0\}) = 0$, and since μ is arbitrary, it follows that $D\mu \leq 0$ a.e. $[\mu]$ on A. But $0 \leq \underline{D}\mu \leq \overline{D}\mu$, so that $D\mu = 0$ a.e. $[\mu]$ on A. $\square$

We are going to show that $D\mu$ exists a.e. $[m]$, and to do this the Lebesgue decomposition theorem is helpful. We write $\mu = \mu_{ac} + \mu_s$, where $\mu_{ac} \ll m$ and $\mu_s \perp m$. If $\mu_s(\mathbb{R}) = 0$, then by 2.3.7, $D\mu = D\mu_{ac}$ a.e. $[m]$, and we are done. Hence we may assume that $\mu_s(\mathbb{R}) > 0$. Then $\mu_{ac}(\mathbb{R}) < \mu(\mathbb{R})$. Since $\mu_s \perp m$, there exists a Borel set S such that $m(S) = 0$ and $\mu_s(S^c) = 0$. Thus $\mu_{ac}(S^c) = 0$ as well. Hence $\mu_{ac} \perp \mu_s$.

By the Radon–Nikodym theorem, we have $\mu_{ac}(F) = \int_F g\,dm$ where $F \subseteq \mathbb{R}$ is Borel and g is a Borel measurable function g. As might be expected, only g is the derivative of μ, hence $D\mu = g$ a.e. $[m]$.

2.3.9 **Theorem.** Let μ be a positive finite measure on $\mathcal{B}(\mathbb{R})$ that is finite on bounded sets, and let $\nu = \mu - \int_0^x g\,dm$, where $g(x) = \mu([0,x])$ if $x \geq 0$ and $g(x) = -\mu([x,0))$ if $x < 0$. Then $D\mu$ exists a.e. $[m]$, and coincides a.e. $[m]$ with the Radon–Nikodym derivative $f = d\mu/dm$.

Proof. If $a < b$ and P is a partition of $[a,b]$ with $a = x_0 < x_1 < \cdots < x_n = b$, then

$$\mu_{ac}(b) - \mu_{ac}(a) = \int_a^b f(x)\,dm = \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} f(x)\,dm.$$

If $x \in (x_{j-1}, x_j)$ for each j, and $\Delta x = \max_j (x_j - x_{j-1})$, then $|f(x) - f(x_j)| < \varepsilon$ for $x \in E$, where $E \subseteq [a,b]$ and $m(E) = b - a$. Thus $|f(x) - f(x_j)| < \varepsilon$ a.e. $[m]$ on $[x_{j-1}, x_j]$, and the Radon–Nikodym derivative $f = d\mu/dm$ satisfies

$$\frac{\mu_{ac}(x_j) - \mu_{ac}(x_{j-1})}{m([x_{j-1}, x_j])} < f(x_j) + \varepsilon.$$

Hence $Df \leq f$ a.e. $[m]$ on $[a,b]$. Therefore, if $E_\varepsilon = \{x \in (a,b) : Df(x) > f(x) + \varepsilon\}$, then $m(E_\varepsilon) = 0$. Since $\bigcup_{\varepsilon \in \mathbb{Q}^+} E_\varepsilon = \{x : Df(x) > f(x)\}$, we have $Df \leq f$ a.e. $[m]$. Replacing μ_1 by $-\mu_1$ and g by $-g$ we obtain $Df \geq f$ a.e. $[m]$. By 2.3.4(c), g is finite a.e. $[m]$, and the result follows. □

We now return to measures on the real line.

2.3.9 **Theorem.** Let $f : [a,b] \to \mathbb{R}$ be a Borel measurable function. Then the derivative of f exists at almost every point of $[a,b]$ (finite or infinite) with respect to Lebesgue measure. Thus by 2.3.5 a function of bounded variation is differentiable almost everywhere.

Proof. Since f has only countably many discontinuities, we may assume without loss of generality that f takes on finite values at all points. Monotonicity on $[a,b]$ is therefore a circumstance, not a hypothesis.

Let μ be the Lebesgue–Stieltjes measure corresponding to f by 2.3.8. $D\mu$ exists a.e. $[m]$, so it suffices to show that $Df = D\mu$ a.e. $[m]$. If $x \in (a,b)$ and μ is differentiable at x, then f is [illegible] at x by definition if $\lim_{y \to x} [f(y) - f(x)]/(y - x)$ [illegible] x, then there is a sequence $h_n \to 0$

where all A_k of the same sign and $|f(x_k) - f(y_k)| < \varepsilon/n$ for all k. Combining all $A_k > 0$, we can find numbers $\xi_k > 0$ such that

$$\frac{f(x_k) - f(y_k)}{x_k - y_k} = \frac{\varepsilon}{n}$$

for $x_k < y_k$ and either y_k or x_k arbitrarily close to x. We may assume that f decreases on $[x, x_k]$ and $[y_k, x]$. Then we would have the ratio $(f(x_k) - f(y_k))/(x_k - y_k) \le \varepsilon/n$, a contradiction. □

We now prove the main theorem on absolutely continuous functions.

2.3.10 Theorem Let f be absolutely continuous on $[a, b]$ with $f(a) = \int_a^a g(t)\,dt$, as in 2.3.4. Then $f' = g$ almost everywhere on $[a, b]$ (Lebesgue measure). Thus by 2.3.4, f is absolutely continuous iff f is an integral of its derivative (or is).

$$f(x) - f(a) = \int_a^x f'(t)\,dt, \qquad a \le x \le b$$

Proof We may assume $g \ge 0$. (If not, consider g^+ and g^-.) If $\mu(A) = \int_A g\,dm$, $A \in \mathcal{B}([a,b])$, then $D\mu_f = g$ a.e. $[m]$ by 2.3.8. But if $a \le x \le y \le b$, then $\mu([x, y]) = f(y) - f(x)$, so that μ is the Lebesgue–Stieltjes measure corresponding to f. Thus by the proof of 2.3.9, $D\mu(x) = f'(x)$ a.e. $[m]$. □

Problems

1. Let F be a bounded distribution function on $\mathbb{R}$. Use the Lebesgue decomposition theorem to show that F may be represented uniquely (up to additive constants) as $F_1 + F_2 + F_3$, where the distribution functions F_1, F_2, F_3 and the corresponding Lebesgue–Stieltjes measures μ_1, μ_2, μ_3 have the following properties.
   (a) F_1 is discrete (that is, μ_1 is concentrated on a countable set of points).
   (b) F_2 is absolutely continuous (μ_2 is absolutely continuous with respect to Lebesgue measure; see 2.3.4).
   (c) F_3 is continuous and singular (that is, μ_3 is singular with respect to Lebesgue measure).
2. If f is an increasing function from $[a, b]$ to $\mathbb{R}$, show that $\int_a^b f'(x)\,dx \le f(b) - f(a)$. (The inequality may be strict, as Problem 1 shows. Note that by 2.3.9, f' exists a.e., but sometimes f' may be defined arbitrarily on an exceptional set of Lebesgue measure 0.)
3. (The Cantor function.) Let $A_1, A_2, \ldots$ be the sets removed from $[0, 1]$ to form the Cantor ternary set (see Problem 7, section 1.1). Define functions

$F_{n,r}$ $[0,1]$ r $[0,1]$ as follows. Let $\{r_{n,1},\dots,r_{n,2^n}\}$ be the set of numbers of the form $F_{n,r}$ arranged in increasing order. For example, if $n=2$,

$$\{r_{2,1},\dots,r_{2,4}\} = \{(0,0),(0,1),(1,0),(1,1)\}$$
$$= \{(0,0),(0,\tfrac{1}{2}),(\tfrac{1}{2},0),(\tfrac{1}{2},\tfrac{1}{2})\}$$
$$= \{0,\tfrac{1}{3},\tfrac{2}{3},1\}.$$

Define

$$F_{n,r}(0) = 0,$$
$$F_{n,r}(x) = F_{n,r}(r_{n,k}) \quad \text{if} \quad x \in [r_{n,k}, r_{n,k+1}), \quad k = 1,\dots,2^n - 1,$$
$$F_{n,r}(1) = 1.$$

Complete the specification of $F_{n,r}$ by interpolating linearly. For $n=2$ and $r=\frac{2}{3}$, $F_{2,2/3}$ is the case.

$$\{r_{2,1},\dots,r_{2,4}\} = \{(0,0),(0,1),(1,0),(1,1)\}$$
$$= \{0,\tfrac{1}{3},\tfrac{2}{3},1\}.$$

Show that $F_{n,r}(x) \to F(x)$ for each x, where F is the Cantor function, has the following properties.

(a) F is continuous and increasing.
(b) $F' = 0$ a.e. (not everywhere) [Lebesgue measure].
(c) F is not absolutely continuous.

Figure 2.11 Approximation to the Cantor function

In fact

(b) F is singular—that is, the corresponding Lebesgue-Stieltjes measure μ_F is singular with respect to Lebesgue measure.

6. Let f be an integrable real-valued function on $\mathbb{R}^k$ for an m-dimensional k. If $\mu(E) = \int_E f(x)\,dx$, $E \in \mathfrak{M}(\mathbb{R}^k)$, we know that $D\mu = f$ a.e. (Lebesgue measure). If $D\mu = f$ at x_0, then x_0 is an upper and lower continuity point of f. We have $\mu(C)/m(C) = f(x_0)$ if $x_0 \in C$.

$$\frac{1}{m(C)} \int_C |f(x) - f(x_0)|\,dx \to 0 \quad \text{as} \quad \text{diam } C \to 0.$$

In fact, show that

$$\frac{1}{m(C)} \int_C |f(x) - f(x_0)|\,dx \to 0 \quad \text{as} \quad \text{diam } C \to 0$$

for almost every x_0. The set of points x_0 is called the Lebesgue set of f.

8. The problem relates to various concepts discussed in Section 2.3. In all cases, f is a real-valued function defined on the closed bounded interval $[a, b]$. Establish the following:

(a) If f is continuous, f need not be of bounded variation.

(c) If f is continuous and increasing (more generally, monotone), f need not be absolutely continuous.

(b) If f satisfies a Lipschitz condition, that is, $|f(x_1) - f(x_2)| \le K|x_1 - x_2|$ for some fixed positive number K and all $x_1, x_2 \in [a, b]$, then f is absolutely continuous.

(d) If f' exists everywhere and is bounded, f is absolutely continuous. (It can also be shown that if f' exists everywhere and is Lebesgue integrable on $[a, b]$, then f is absolutely continuous; see Titchmarsh, 1939, p. 368.)

(e) If f is continuous and f' exists everywhere, f need not be absolutely continuous [consider $f(x) = x^2 \sin(\pi/x)$, $0 < x \le 1$, $f(0) = 0$].

9. The following problem considers the change of variable formula in multiple integrals. Let g be a one-to-one mapping of an open set W onto V. Assume W and V are subsets of $\mathbb{R}^k$. T is assumed not to be one-to-one on all of W, but W can be divided into finitely many disjoint sets $W_1, W_2, \ldots, W_n$ such that T is one-to-one on each W_i. Thus T has a continuously differentiable inverse on each W_i (by the inverse function theorem of advanced calculus; see, for example, Apostol, 1957, p. 144). It also follows from elementary advanced calculus results that for all $x \in V$,

$$\frac{1}{|J|}\left[\text{Vol}\left(\bigcup_{i=1}^{n} T(W_i) \cap S(x, r)\right)\right] \to 0 \quad \text{as} \quad r \to 0,$$

value $\alpha(p)$ is the linear transformation on $\mathbb{R}^n$ represented by the Jacobian matrix of f, evaluated at p. [See Apostol (1957, p. 118).]

(c) Let ϕ be a nonsingular linear transformation on $\mathbb{R}^n$ and define a measure λ on $\mathscr{B}(\mathbb{R}^n)$ by $\lambda(E) = m(\phi[E])$, where m is Lebesgue measure. Show that $\lambda = c\mu$ for some constant $c(\phi)$, and in fact $c(\phi)$ is the absolute value of the determinant of ϕ. [Use translation-invariance of Lebesgue measure (Theorem 4.3, Exercise 3) and the fact that any matrix can be represented as a product of matrices corresponding to elementary row operations.]

Now define a measure μ on $(\mathbb{R}^n, \mathscr{B}^n)$ by $\mu(E) = m(T(E))$. By continuity of T, $T^{-1}(x - \mathbb{R}^n) \in \mathscr{B}^n$, and T^{-1} is a one-to-one linear map of $\mathbb{R}^n$ onto $\mathbb{R}^n$, so $T(E)$ is a Borel set whenever E is. Therefore μ is a well-defined measure. Moreover, μ is finite on compact sets, since T is continuous and hence maps compact sets to compact sets. Thus μ is a regular Borel measure (Theorem 4.8, Exercise 3) and therefore $\mu = \lambda$ on $\mathscr{B}^n$.

Our objective is to show that μ is also translation-invariant and $\mu(D_1 \times \cdots \times D_n) = |D_1| \cdots |D_n|$ for every $D_j \in \mathscr{B}^1$, where $|D_j|$ denotes the length of D_j. [Use the translation-invariance of m to show that $\mu(E + y) = \mu(E)$ for every $y \in \mathbb{R}^n$, where $E + y = \{x + y : x \in E\}$. Then apply Exercise 4.]

(b) Show that it suffices to prove that if $D_1, \ldots, D_n$ are bounded intervals, then $\mu(D_1 \times \cdots \times D_n) = |D_1| \cdots |D_n|$.

(c) Show that μ may be assumed without loss of generality that A is a nonsingular linear transformation. [Hint: $\mu(0) = 0$.]

Now given $\varepsilon > 0$, choose $\alpha \in (0, \frac{1}{2})$ such that

$$(1 - \alpha)^n - \alpha n 2^{n-1} > (1 - 2\alpha)^n > 1 - \varepsilon.$$

Under the assumptions of (b) and (c), by Eq. (*), there is a $\delta > 0$ such that if $|x| < \delta$ then $|Tx| - |x| < \alpha |x| / \sqrt{n}$.

(d) If C is an open cube containing 0 of edge length β and diameter $\sqrt{n}\beta < \delta$, take C_1, C_2 to be open cubes concentric with C, with edge lengths $\beta_1 = (1 - \alpha)\beta$ and $\beta_2 = (1 - 2\alpha)\beta$, respectively. Then the following

(i) if $x \in \mathbb{R}^n$, then $Tx \in C$;

(ii) if y belongs to the boundary of C_1, then $Ty \in C_2$;

(iii) if z is the center of C_1, then $Tz \in C_1$;

(iv) $C_1 - Tz \subset C_2 - Tx$.

To see why (iv) is true, note that $C_1 \subset T[C]$ and $C_2 \subset T[\cdots]$, and compare the edge lengths.

(e) If λ is any measure on $(\mathbb{R}^n, \mathscr{B}^n)$ and $\lambda(0) < \infty$, we say that λ is absolutely continuous with respect to Lebesgue measure. It then follows from Theorem 3.13 that

$$\lambda(T[E]) = \int_E J(x)\, dx \qquad (E \in \mathscr{B}^n),$$

[If this is false, find a compact set K and positive integers n and j such that $m(K) - \int_K g_n(x)\,dx < 2^{-j}$ and K is a finite union of open cubes each of which has measure less than 2^{-j}. Essentially, the idea is to cover K by such cubes and conclude that $m(K) = \int_K g_n$, a contradiction.]

10. If f is a real-valued Borel measurable function on $\mathbb{R}^n$, show that

$$\int_{\mathbb{R}^n} f(x)\,dx = \int_0^\infty m(\{x : f(x) > t\})\,dt$$

in the sense that if one of the two integrals exists (and is finite), then so does the other, and the two integrals are equal.

7. (Baire's differentiation theorem) Let $f_1, f_2, \ldots$ be continuous functions on $\mathbb{R}^n$, and assume that for each $x \in \mathbb{R}^n$, $\sum_{j=1}^\infty f_j(x)$ converges to a finite number $f(x)$. Show that $\sum_{j=1}^\infty f_j(x) = f(x)$ almost everywhere (Lebesgue measure).

Outline:

(a) It suffices to restrict the domain of all functions to $[0,1]^n$ and to assume all functions are nonnegative. Use Fatou's lemma to show that $\sum_{j=1}^N f_j(x) < C$ for some N and some $C < \infty$.

(b) Choose $\varepsilon_1, \varepsilon_2, \ldots$ such that $\sum_{j=1}^\infty \varepsilon_j(1) \le 2^{-k}$ for $k = 1, 2, \ldots$, and apply part (a) to the function $g = f(x) - \sum_{j=1}^N f_j(x) - \sum_{j=1}^\infty \varepsilon_j(x)$.

2.4 L^p Spaces

If $(\Omega, \mathcal{F}, \mu)$ is a measure space and p is a real number with $p \ge 1$, the set of all Borel measurable functions f such that $\int |f|^p$ is nondegenerate, we merely enumerate properties. It is crucial to fully develop these properties, as we will be constantly working with complex-valued functions.

2.47 Definition. Let $(\Omega, \mathcal{F})$ be a measurable space and let f be a complex-valued function on Ω, so that $f = \operatorname{Re} f + i \operatorname{Im} f$. We say that f is a complex-valued Borel measurable function on $(\Omega, \mathcal{F})$ if and only if $\operatorname{Re} f$ and $\operatorname{Im} f$ are real-valued Borel measurable functions. If μ is a measure on $(\Omega, \mathcal{F})$, we define

$$\int_\Omega f\,d\mu = \int_\Omega \operatorname{Re} f(\omega)\,d\mu + i \int_\Omega \operatorname{Im} f(\omega)\,d\mu,$$

provided $\int_\Omega \operatorname{Re} f\,d\mu$ and $\int_\Omega \operatorname{Im} f\,d\mu$ are both finite. In this case, we say that f is integrable. Since in working with complex-valued functions we do not consider any cases in which integrals exist but are not finite.

The following result was established earlier for real-valued f [see 2.5 (iv)]; it is still valid in the complex case, but the proof must be modified.

3.4.2 *Lemma.* If f is an integrable

$$\left|\int_E f\,d\mu\right| \le \int_E |f|\,d\mu.$$

Proof. If $\int_E f\,d\mu = re^{i\theta}$, $r \ge 0$, then $\int_E e^{-i\theta} f\,d\mu = r = \left|\int_E f\,d\mu\right|$. But if $f(x) = \rho(x)e^{i\theta(x)}$, letting $\rho \ge 0$ we have

$$\begin{aligned}\int_E e^{-i\theta} f\,d\mu &= \int_E e^{-i\theta}\rho e^{i\theta}\,d\mu \\ &= \int_E \rho e^{i(\theta-\theta)}\,d\mu = \int_E \rho\,d\mu \quad \text{since } \rho \text{ is real} \\ &\le \int_E \rho\,d\mu = \int_E |f|\,d\mu. \quad \square\end{aligned}$$

Many other standard properties of the integral carry over to the complex case, in particular 3.3.3(b), 3.3.4, 3.3.5, 3.3.6, 3.3.7, 3.3.8, 3.3.9(b) and (c), 3.3.10, 3.3.11, 3.3.12, 3.3.13 and 3.3.14. In almost all cases the result is an immediate consequence of the fact that in applying a complex-valued function to integrating the real and imaginary parts separately. Only two theorems require additional comment. To prove that if f is integrable on E so is $|f|$, note that $\operatorname{Re} f \le |f|$ implies $\operatorname{Re} f = |\operatorname{Re} f| \le |f|$. Then apply the proof of the theorem as $f_n = \operatorname{Re} f$ and note that $|f_n - f| \le |f_n| + |f| \le 2|f|$.

If $p > 0$ we define the space $L^p = L^p(\Omega, \mathcal{F}, \mu)$ as the collection of all complex valued Borel measurable functions f such that $\int_\Omega |f|^p\,d\mu < \infty$. We set

$$\|f\|_p = \left(\int_\Omega |f|^p\,d\mu\right)^{1/p}, \qquad f \in L^p.$$

It follows that for any complex number α, $\|\alpha f\|_p = |\alpha|\,\|f\|_p$, $f \in L^p$.

We are going to show that L^p forms a linear space over the complex field. The key steps in the proof are the Hölder and Minkowski inequalities, which we now develop.

3.4.3 *Lemma.* If $a, b > 0$, $p > 0$, $q > 0$, $\frac{1}{p} + \frac{1}{q} = 1$, then $a^{1/p} b^{1/q} \le \frac{a}{p} + \frac{b}{q}$.

Proof. This statement to be proved is equivalent to $\log a^{1/p} b^{1/q} \le \log\left(\frac{a}{p} + \frac{b}{q}\right)$, or $\frac{1}{p}\log a + \frac{1}{q}\log b \le \log\left(\frac{a}{p} + \frac{b}{q}\right)$, which holds because $\log x$ is concave. If φ is a nonnegative convex function on the interval $I \subseteq R$ that is continuous on I, we say φ is *convex* if $\varphi(px + qy) \le p\varphi(x) + q\varphi(y)$, $x, y \in I$, $p, q > 0$, $p + q = 1$. This is

Thus, assume $x < y$ and write

$$\varphi(x) + \beta(y - x) = \varphi(x) + \beta(y - x) - \beta(0) = \varphi(x) - \beta(0) + \beta(y - x) = \varphi(x) - \beta(x - x) + \beta(y - x) \geq \varphi(x) - \varphi(x) + \varphi(y) = \varphi(y) \quad \text{for some} \quad \beta \in \mathbb{R}$$

with $x \leq \gamma \leq y$. Or, $\varphi(x) + \beta(y - x) \geq \varphi(y)$ if $\beta \geq \varphi'(\gamma)$. Now $\varphi'(x) = \sigma(x)$ is increasing on I.

2.4.4 *Corollary.* If $\alpha, \beta > 0$, $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, and $\alpha^p = \beta^q$, then $\alpha\beta \leq \frac{1}{p}\alpha^p + \frac{1}{q}\beta^q$.

Proof. Let $\varphi(x) = e^x$ in 2.4.3. Set $x = \log(\alpha^p)$, $y = \log(\beta^q)$, $\alpha = e^{x/p}$, $\beta = e^{y/q}$. □

2.4.5 *Hölder Inequality.* Let $(X, \mathcal{M}, \mu)$ be a measure space, let p and q be positive numbers, $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in L^p$ and $g \in L^q$, then $fg \in L^1$ and $\|fg\|_1 \leq \|f\|_p \|g\|_q$.

Proof. If $\|f\|_p = 0$ or $\|g\|_q = 0$, then $f = 0$ a.e. or $g = 0$ a.e., so that $fg = 0$ a.e. and the inequality is immediate. Thus we may assume $\|f\|_p > 0$ and $\|g\|_q > 0$. Then

$$\frac{|f(x)g(x)|}{\|f\|_p \|g\|_q} \leq \frac{|f(x)|^p}{p\|f\|_p^p} + \frac{|g(x)|^q}{q\|g\|_q^q}$$

and integration yields

$$\frac{\int_X |fg|\,d\mu}{\|f\|_p \|g\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1. \qquad \square$$

When $p = q = 2$ we obtain

$$\int_X |fg|\,d\mu \leq \left(\int_X |f|^2\,d\mu \int_X |g|^2\,d\mu\right)^{1/2}$$

and thus, using 2.4.2, we have the *Cauchy–Schwarz inequality*: If f and $g \in L^2$, then $fg \in L^1$ and

$$\left|\int_X f\bar{g}\,d\mu\right| \leq \left(\int_X |f|^2\,d\mu \int_X |g|^2\,d\mu\right)^{1/2}.$$

Where $\bar{g}$ is the complex conjugate of g. (The reason for replacing g by $\bar{g}$ is to make the inner product agree with the Hilbert space result to be presented in Chapter 11.)

3.4.6 *Lemma.* If $\alpha > \beta \geq 0$, $\gamma \geq 1$, then $(\alpha - \beta)^\gamma \leq 2^{\gamma-1}(\alpha^\gamma - \beta^\gamma)$.

Proof. Let $\varphi(x) = x^\gamma - (x-1)^\gamma - 2^{\gamma-1}(x^\gamma - 1)$ for $x \geq 1$. Then $\varphi'(x) = \gamma x^{\gamma-1} - \gamma(x-1)^{\gamma-1} - 2^{\gamma-1}\gamma x^{\gamma-1} \leq 0$, since $x - 1 \leq x$. Thus

$$\varphi(x) \geq \varphi(1) = 0 \quad \text{for} \quad 1 \leq x \leq 2\alpha, \quad \text{that is,} \quad x \leq \alpha,$$
$$\varphi(x) = 0 \quad \text{at} \quad x = \alpha,$$
$$\varphi(x) \to 0 \quad \text{for} \quad x \to \infty.$$

The maximum therefore occurs at $x = \alpha$. Hence

$$\alpha^\gamma - (\alpha-\beta)^\gamma - 2^{\gamma-1}(\alpha^\gamma - \beta^\gamma) \geq \beta^\gamma - (\alpha-\beta)^\gamma - 2^{\gamma-1}(\alpha^\gamma - \beta^\gamma) = \varphi(\alpha) \geq 0. \quad \square$$

3.4.7 *Minkowski inequality.* If $f, g \in L^p(X, \mathcal{S}, \mu)$, $1 \leq p < \infty$, then $f + g \in L^p$ and $\|f+g\|_p \leq \|f\|_p + \|g\|_p$.

Proof. By 3.4.6, $|f-g|^p \leq (|f| + |g|)^p \leq 2^{p-1}(|f|^p + |g|^p)$; hence if $f, g \in L^p$ implies $f - g \in L^p$. Now the inequality is clear when $p = 1$, so assume $p > 1$ and choose q such that $(1/p) + (1/q) = 1$. Then

$$|f+g|^p = |f+g||f+g|^{p-1} \leq |f||f+g|^{p-1} + |g||f+g|^{p-1} \tag{1}$$

Now $(f+g)^{p-1} \in L^q$; for

$$\int |f+g|^{(p-1)q} = \int |f+g|^p = \int |f+g|^{p-1}|f+g| < \infty,$$

hence

$$\int_{[0]} |f+g|^{(p-1)q} d\mu = \int_{[0]} |f+g|^p d\mu < \infty.$$

Since f and g belong to L^p and $|f+g|^{p-1} \in L^q$, Hölder's inequality implies that $|f||f+g|^{p-1}$ and $|g||f+g|^{p-1}$ belong to L^1, and

$$\int_{[0]} |f||f+g|^{p-1} d\mu + \int_{[0]} |g||f+g|^{p-1} d\mu \leq \left[\int_{[0]} |f+g|^{(p-1)q} d\mu\right]^{1/q} \cdot \left[\|f\|_p + \|g\|_p\right]$$
$$= \|f+g\|_p^{p/q} \cdot \left[\|f\|_p + \|g\|_p\right]. \tag{2}$$

$$\int_{[0]} |g||f+g|^{p-1} d\mu \leq \|g\|_p \cdot \|f+g\|_p^{p/q} \tag{3}$$

By Eqs. (1), $\|f+g\|_p^p \leq \|f\|_p + \|g\|_p \cdot \|f+g\|_p^{p/q}$. Since $p - (p/q) = 1$, the result follows. $\square$

By Minkowski's inequality and the fact that $a|f| = |af|$ for $f \in L$, $a \in \mathbb{C}$, $\|f\| \geq 0$ and $\|af\| = |a|\,\|f\|$. Moreover, since $\mathbb{C}$ is a vector space over the complex field, Problem 4.5.5 shows that there is a natural norm on this space in L^* by virtue of the fact that $|\cdot|$ is a norm on $\mathbb{C}$.

4.4.8 Definitions and Comments. A seminorm on a vector space L over the real or complex field is a real valued function $\|\cdot\|$ on L with the following properties:

$$\|f\| \geq 0,$$

$$a\|f\| = |a|\,\|f\| \quad \text{for each scalar } a,$$

consequently, if $f = 0$, then $\|f\| = 0$.

$$\|f + g\| \leq \|f\| + \|g\|$$

(f and g are arbitrary elements of L.) $\|\cdot\|$ is a norm if it has the additional property that $\|f\| = 0$ implies $f = 0$. $\|\cdot\|$ is said to be a norm.

Now $\|\cdot\|_p$ is a seminorm on L^p: the first two properties follow from the definition of $\|\cdot\|_p$, and the last property is a consequence of Minkowski's inequality.

We now indicate how to change $\|\cdot\|_p$ from a seminorm to a norm by passing to equivalence classes, as follows.

If $f, g \in \mathcal{P}(\Omega, \mathcal{F}, m)$, define $f \sim g$ if $\|f - g\|_p = 0$ and $|f|_p$. Then $f \sim g$ is the same for all f in a given equivalence class, by Problem 4.5.5. Thus if L^p is the collection of equivalence classes, L^p becomes a linear space, and $\|\cdot\|_p$ is a seminorm on L^p. In fact $\|\cdot\|_p$ is a norm since $\|f\|_p = 0$ implies $f = 0$ a.e., [m], by Problem 4.5.5(b).

If $\|\cdot\|$ is a seminorm on a vector space, we have a natural notion of distance: $d(f, g) = \|f - g\|$. By definition of seminorm we have:

$$d(f, g) \geq 0,$$

$$d(f, g) = 0 \quad \text{if} \quad f = g,$$

$$d(f, g) = d(g, f),$$

$$d(f, h) \leq d(f, g) + d(g, h).$$

These all have all the properties of a metric, except that $d(f, g) = 0$ does not necessarily imply $f = g$. We call d a *pseudometric*. If $\|\cdot\|$ is a norm, d is actually a metric. This is just a convenience of terminology between seminorm and pseudometric, but these terms seem to be most popular, although "pseudonorm" is sometimes used for "seminorm."

One of the first questions that arises in any attempt to solve the problem of completeness is whether or not Cauchy sequences converge. We are going to show that the L^p spaces are complete. The following result will be needed in almost all probability and likely to reappear in immediately various applications of analysis as well.

2.4.9 *Chebyshev's Inequality.* Let f be a nonnegative extended real valued Borel measurable function on $(\Omega, \mathcal{F}, \mu)$. If $0 < p < \infty$ and $0 < \alpha < \infty$ then

$$\mu\{\omega : f(\omega) \geq \alpha\} \leq \frac{1}{\alpha^p} \int_\Omega f^p \, d\mu.$$

The following version is often applied in probability. If χ is an extended real valued Borel measurable function on $(\Omega, \mathcal{F})$ and P is a probability measure on $\mathcal{F}$, then

$$m = \int_\Omega \chi \, dP \qquad \text{assumed finite, so that } \chi \text{ is finite a.e. } [P],$$

$$\sigma^2 = \int_\Omega (\chi - m)^2 \, dP$$

If $0 < k < \infty$ then

$$P\{\omega : |\chi(\omega) - m| \geq k\sigma\} \leq \frac{1}{k^2}.$$

This follows from the first version with $f = (\chi - m)^2$, $\alpha = k^2\sigma^2$, $p = 2$.

Proof.

$$\int_\Omega f^p \, d\mu = \int_{\{\omega : f(\omega) \geq \alpha\}} f^p \, d\mu + \int_{\{\omega : f(\omega) < \alpha\}} f^p \, d\mu \geq \alpha^p \mu\{\omega : f(\omega) \geq \alpha\}. \quad \square$$

One more result very much will be needed.

2.4.10 *Lemma.* If $g_1, g_2, \ldots \in L^p(\mu)$, $0 < p < \infty$, and $\|g_k - g_{k+1}\|_p \leq \left(\frac{1}{2}\right)^k$, $k = 1, 2, \ldots$, then g_k converges a.e.

Proof. Let $A_k = \{\omega : |g_k(\omega) - g_{k+1}(\omega)| \geq 2^{-k}\}$. Then by 2.4.9,

$$\mu(A_k) \leq 2^{kp} \|g_k - g_{k+1}\|_p^p \leq 2^{-kp}.$$

By [illegible] if $\omega \notin \limsup A_k$, then $g_k(\omega) = g_{k+1}(\omega)$ for all $k \geq k_0$ for some k_0, so $\{g_k(\omega)\}$ is a Cauchy sequence of complex numbers, and therefore converges. $\square$

Now, the main result

9.1.13 Completeness of L^p, $1 \le p < \infty$. If $f_1, f_2, \ldots$ form a Cauchy sequence in L^p, that is, $\|f_n - f_m\|_p \to 0$ as $n, m \to \infty$, then there is an $f \in L^p$ such that $\|f_n - f\|_p \to 0$.

Proof. Take a subsequence $\{f_{n_k}\}$ with $\|f_{n_{k+1}} - f_{n_k}\|_p < 2^{-k}$ for $n_{k+1} \ge n_k$, and let $g_k = f_{n_k}$. In general, having chosen $g_1, \ldots, g_j$ and $n_1 < \cdots < n_j$, let $n_{j+1} > n_j$ be such that $\|f_n - f_m\|_p < (2^j)^{-1}$ for $n, m \ge n_{j+1}$, and let $g_{j+1} = f_{n_{j+1}}$. Then g_j converges a.e. to a limit function f.

Given $\varepsilon > 0$, choose N such that $\|f_n - f_m\|_p^p < \varepsilon$ for $n, m \ge N$. Fix $n \ge N$ and let $m \to \infty$ through values in the subsequence, that is, let $n = n_0, n_1, \ldots$. Then

$$\begin{aligned} \varepsilon &\ge \liminf_{m \to \infty} \|f_n - f_m\|_p^p = \liminf_{m \to \infty} \int_\Omega |f_n - f_m|^p \, d\mu \\ &\ge \int_\Omega \liminf_{m \to \infty} |f_n - f_m|^p \, d\mu \qquad \text{by Fatou's lemma} \\ &= \|f_n - f\|_p^p. \end{aligned}$$

Thus $\|f_n - f\|_p \le \varepsilon^{1/p}$. Since $f = f - f_n + f_n$, we have $f \in L^p$.

9.1.14 Examples and comments. Let us be more explicit in Example 9.1.12. Let Ω be a set, $\mathcal{A}$ a σ-algebra of subsets of Ω, and let μ be a countably additive measure. A real-valued function on Ω may be represented as a sequence $(a_1, a_2, \ldots)$ of real numbers, provided we write $f = (a_n, n = 1, 2, \ldots)$. An integral on this space is really a sum. See Problem 9.1.1 of

$$\int_\Omega f \, d\mu = \sum_{n=1}^{\infty} a_n \mu_n$$

where the series is interpreted as $\sum_{n=1}^{\infty} a_n^+ \mu_n - \sum_{n=1}^{\infty} a_n^- \mu_n$. This is one of the forms $\infty - \infty$ (if f is not integrable) does not exist. Three interesting cases occur.

(1) $\sum_{n=1}^{\infty} a_n^+ = \infty$, $\sum_{n=1}^{\infty} a_n^- < \infty$. The series diverges to $+\infty$ and the integral is $+\infty$.

(2) $\sum_{n=1}^{\infty} a_n = +\infty$, $\sum_{n=1}^{\infty} a_n = -\infty$. The series diverges to $+\infty$ and the integral is $+\infty$.

(3) $\sum_{n=1}^{\infty} a_n^+ < \infty$, $\sum_{n=1}^{\infty} a_n^- = \infty$. The series is absolutely convergent and the integral equals the sum of the series.

(4) $\sum_{n=1}^{\infty} a_n^+ = \infty$, $\sum_{n=1}^{\infty} a_n^- = \infty$. The series may not be absolutely convergent; it may or may not converge conditionally. Whether it does or not, the integral does not exist. This is when summation is considered from the point

of view of Lebesgue integration theory, sense that convergence unconditionally but not absolutely are ignored.

If μ is counting measure on the positive integers, so that $\mu(\{n\}) = 1$ for each n, we recover the case of counting measure. In some analysis classes we write

$$\int_{[0,\infty)} f\,d\mu = \sum_{n=1}^{\infty} f(n),$$

where the series is interpreted as $\sum_{n=1}^{\infty} f(n) = \lim_{N\to\infty} \sum_{n=1}^{N} f(n)$.

If $\Gamma = \{z_n : n = 1, 2, \ldots\}$ is a sequence of complex numbers and μ is counting measure,

$$\int_{\Gamma} f\,d\mu = \sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \operatorname{Re}(a_n) + i \sum_{n=1}^{\infty} \operatorname{Im}(a_n),$$

the integral is defined provided $\sum_{n=1}^{\infty} |a_n| < \infty$.

Now let Ω be an arbitrary set, and take $\mathcal{F}$ to be all subsets of Ω and μ to be counting measure. If $f = \{f(\omega) : \omega \in \Omega\}$ is a nonnegative real valued function on Ω, then [Problem 1.6.7]

$$\int_{\Omega} f\,d\mu = \sum_{\omega\in\Omega} f(\omega), \tag{1}$$

where the series is defined as $\sup\{\sum_{\omega\in E} f(\omega) : E \subset \Omega, E \text{ finite}\}$. If $f(\omega) > 0$ for uncountably many ω, then for some $\delta > 0$ we have $f(\omega) > \delta$ for infinitely many ω so the $\sum_{\omega} f(\omega) = \infty$.

If the nonnegativity hypothesis is dropped, we apply the above results to f^+ and f^- to again obtain Eq. (1), where the series is interpreted as $\sum_{\omega} f^+(\omega) - \sum_{\omega} f^-(\omega)$. If f is complex-valued, Eq. (1) applies with the series interpreted as $\sum_{\omega} \operatorname{Re}(f(\omega)) + i \sum_{\omega} \operatorname{Im}(f(\omega))$; the integral is finite provided $\sum_{\omega} |f(\omega)| < \infty$.

The space $L^1(\Omega, \mathcal{F}, \mu)$ will be denoted by $l^1(\Omega)$; it consists of all complex-valued functions $f(\omega) : \omega \in \Omega$ such that $f(\omega) = 0$ for all but countably many ω and

$$\|f\|_1 = \sum_{\omega\in\Omega} |f(\omega)| < \infty.$$

If Ω is the set of positive integers, the space $l^1(\Omega)$ will be denoted simply by l^1; it consists of all sequences $f = \{f(n)\}$ of complex numbers such that

$$\|f\|_1 = \sum_{n=1}^{\infty} |f(n)| < \infty.$$

It will be useful to state the Hölder and Minkowski inequalities for sums. If $f=\{f(n)\}$ and $g=\{g(n)\}$, where $1<p<\infty$, $\frac{1}{p}+\frac{1}{q}=1$, then $fg\in l^1(\mathbb{Z})$ and

$$\sum_{n=-\infty}^{\infty}|f(n)g(n)|\le\left(\sum_{n=-\infty}^{\infty}|f(n)|^p\right)^{1/p}\left(\sum_{n=-\infty}^{\infty}|g(n)|^q\right)^{1/q}.$$

If $f,g\in l^p(\mathbb{Z})$, $1\le p<\infty$, then $f+g\in l^p(\mathbb{Z})$ and

$$\left(\sum_{n=-\infty}^{\infty}|f(n)+g(n)|^p\right)^{1/p}\le\left(\sum_{n=-\infty}^{\infty}|f(n)|^p\right)^{1/p}+\left(\sum_{n=-\infty}^{\infty}|g(n)|^p\right)^{1/p}.$$

As in 2.4.5, we obtain the Cauchy–Schwarz inequality as the case $p=2$ of the Hölder inequality. If $f,g\in l^2(\mathbb{Z})$, then $fg\in l^1(\mathbb{Z})$ and

$$\left|\sum_{n=-\infty}^{\infty}f(n)\overline{g(n)}\right|\le\left(\sum_{n=-\infty}^{\infty}|f(n)|^2\right)^{1/2}\left(\sum_{n=-\infty}^{\infty}|g(n)|^2\right)^{1/2}.$$

If in the above discussion we replace $\mathbb{Z}$ by $\{1,2,\dots,n\}$, all assumptions and conclusions are obtainable, and all the spaces $l^p(\mathbb{Z})$ become $\mathbb{C}^n$.

Let $1\le p<\infty$. $\|\cdot\|_p$ is not a seminorm on $L^p(\Omega,\mathcal{B},\mu)$. For let A and B be disjoint sets with $0<\mu(A)<\infty$ and $0<\mu(B)<\infty$. Assume finite and positive. If $f=\chi_A$, $g=\chi_B$, then

$$\|f-g\|_p=\left(\int_{A\cup B}|\chi_A-\chi_B|^p\,d\mu\right)^{1/p}=\left(\int_{A\cup B}1\,d\mu\right)^{1/p}=\mu(A)+\mu(B)>0$$

but $\|f\|_p-\|g\|_p=\mu(A)^{1/p}-\mu(B)^{1/p}=0$. If $\mu(A)+\mu(B)\ne 2$, then $\|f\|_p-\|g\|_p<\|f-g\|_p$. If $\mu(A)+\mu(B)=2$, then $\|f\|_p-\|g\|_p=\|f-g\|_p$. Thus the triangle inequality fails. We can, however, describe convergence in L^p. Theorem 1. In the following sense. We use the inequality

$$|a+b|^p\le 2^p(|a|^p+|b|^p),\qquad a,b\in\mathbb{C},\qquad 0<p<\infty$$

which is proved by considering $(a+b)^p=a^p+b^p$. It follows that

$$\int_\Omega|f+g|^p\,d\mu\le 2^p\int_\Omega|f|^p\,d\mu+2^p\int_\Omega|g|^p\,d\mu,\qquad f,g\in L^p. \tag{2}$$

and we confirm that $d(f,g)=\int_\Omega|f-g|^p\,d\mu$ defines a pseudometric on L^p. In fact the pseudometric is complete (every Cauchy sequence converges); for Eq. (2)

implies that $\alpha f, f+g \in \mathcal{L}^1$, that $\int \alpha f = \alpha \int f$, and that the triangle inequality $\int |f+g| \le \int |f| + \int |g|$ holds.

If $\mathcal{M}$ is an interval of reals, $\mathcal{F}$ is the class of Borel sets of $\mathcal{M}$, and μ is Lebesgue measure, the space $\mathcal{L}^1(\mathcal{M}, \mathcal{F}, \mu)$ will be denoted by $L^1(\mathcal{M})$. Thus, for example, $L^1[a,b]$ is the set of all complex-valued Borel measurable functions f on $[a,b]$ such that

$$\|f\|_1 = \int_a^b |f(x)|\,dx < \infty.$$

If f is a complex-valued Borel measurable function on $(\Omega, \mathcal{F}, \mu)$ and $f_1, f_2, \ldots \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$, we say that the sequence $\{f_n\}$ converges to f in L^1 if $\|f_n - f\|_1 \to 0$, that is, if $\int |f_n - f| \to 0$ as $n \to \infty$. We use the notation $f_n \xrightarrow{L^1} f$. In Section 2.5 we shall compare various types of convergence of sequences of measurable functions. We now show that every $f \in L^1$ is an L^1-limit of simple functions.

2.4.12 *Theorem.* Let $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$, $0 < \mu(\Omega) < \infty$. Then, for each $\varepsilon > 0$, there is a simple function $g \in \mathcal{F}$ such that $\|f - g\|_1 < \varepsilon$. If f is real-valued, g can be chosen to be real-valued; if $f \ge 0$, g can be chosen to be nonnegative; if f is bounded, g can be chosen so that $\sup |g| \le \sup |f|$.

Proof. This follows from 2.3.10 and 2.3.11. □

If we specialize to functions on R^1 and Lebesgue–Stieltjes measures, we may obtain another useful approximation theorem.

2.4.13 *Theorem.* Let $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$, $0 < \mu(\Omega) < \infty$, where $\Omega = \mathbb{R}^1$ or $\Omega = \mathbb{R}^1$, and μ is a Lebesgue–Stieltjes measure. If $\varepsilon > 0$, there is a continuous function $g \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ such that $\int |f - g| < \varepsilon$. Furthermore, g can be chosen so that $\sup |g| \le \sup |f|$. Thus the conclusion holds for bounded f.

Proof. By 2.4.12, it is sufficient to show that an indicator I_A in L^1 can be approximated in the L^1 sense by a continuous function with absolute value at most 1. Now $I_A \in L^1$ means that $\mu(A) < \infty$ by 1.4.1. There is a closed set $C \subset A$ and an open set $V \supset A$ such that $\mu(V - C) < \varepsilon$ by 1.4.7. Let g be a continuous map of $\mathbb{R}^1$ into $[0,1]$ with $g = 1$ on C and $g = 0$ on V^c (exists by Urysohn's lemma). Then

$$\int_\Omega |I_A - g|\,d\mu = \int_{V-C} |I_A - g|\,d\mu$$

But $|f_k - g| \to 0$ on $X - E$ and $|f_k - g| \le 2$, hence

$$\int_X |f_k - g|^p \, d\mu \to 0.$$

Since $\varphi - g = (\varphi - f_k) + (f_k - g)$, we have $\varphi \in L^p(\mu)$.

If $f_1, f_2, f_3, \ldots$ are continuous and f_k converges to f in L^p, it does not follow that f is continuous (see Problem 3).

2.4.13 The Space L^∞. If we wish to define L^p spaces for $p = \infty$, we run into problems immediately. We define the essential supremum of a real-valued Borel measurable function f on $(X, \mathcal{M}, \mu)$ as

$$\operatorname{ess\,sup} f = \inf\{a : \mu(\{x : f(x) > a\}) = 0\},$$

that is, the smallest number a such that $f \le a$ almost everywhere.

If f is a complex valued Borel measurable function on $(X, \mathcal{M}, \mu)$, we define

$$\|f\|_\infty = \operatorname{ess\,sup} |f|.$$

The space $L^\infty(X, \mathcal{M}, \mu)$ is the collection of all f such that $\|f\|_\infty < \infty$. Thus $f \in L^\infty$ iff f is essentially bounded, that is, bounded outside a set of measure 0.

Now $|f - g| \le |f| + |g| \le \|f\|_\infty + \|g\|_\infty$ a.e., hence

$$\|f + g\|_\infty \le \|f\|_\infty + \|g\|_\infty.$$

It is also clear that $f \in L^\infty$ implies $cf \in L^\infty$. The usual properties of a norm are easily checked. Thus L^∞ is a vector space over the complex field, $\|\cdot\|_\infty$ is a seminorm on L^∞, and becomes a norm if we pass to equivalence classes as before.

If $f_1, f_2, f_3, \ldots \in L^\infty$ and $\|f_k - f\|_\infty \to 0$, we write $f_k \xrightarrow{u} f$, we claim that $\|f_k - f\|_\infty \to 0$. For there is a set $E \in \mathcal{M}$ with $\mu(E) = 0$ such that $f_k \to f$ uniformly on $X - E$.

Assume $\|f_k - f\|_\infty > 0$. Given a positive integer k, $|f_k - f| \le \|f_k - f\|_\infty$ a.e., and consequently $|f_k(x) - f(x)| \le \|f_k - f\|_\infty$ for all x except those in a set A_k of measure 0. Hence $|f_k(x) - f(x)| \le \|f_k - f\|_\infty$ for all x except those in $\bigcup_{k=1}^\infty A_k$, which has measure 0, and $f_k \to f$ uniformly on $X - \bigcup_{k=1}^\infty A_k$. Conversely, assume $\mu(A) = 0$ and $f_k \to f$ uniformly on A^c. Given $\varepsilon > 0$, $|f_k - f| \le \varepsilon$ on A^c for sufficiently large k, so that $\|f_k - f\|_\infty \le \varepsilon$. Thus $\|f_k - f\|_\infty \to 0$ for large enough k, and the result follows.

An identical argument shows that $\{f_k\}$ is a Cauchy sequence in L^∞ iff $\|f_k - f_m\|_\infty \to 0$ as $k, m \to \infty$ iff there is a set $A \in \mathcal{M}$ with $\mu(A) = 0$ and $f_k - f_m \to 0$ uniformly on A^c.

It is sometimes true that the Minkowski inequality still holds when $p = 1$ or $p = \infty$, and we have shown above that the Minkowski inequality holds when $p = \infty$.

To show that f^* is a candidate for $\lim_{n\to\infty} f_n$ in an L^∞ sense, we need to be sure of measurability such that $f_n(x) \to f(x)$ — for merely for $x \in A^c$. But that $f_n(x)$ converges to a finite limit for each $x \in A^c$ and the convergence is uniform on A^c. If we define $f(x) = 0$ for $x \in A$, we have $f \in L^\infty$ and $f_n \xrightarrow{L^\infty} f$.

Theorem 2.4.13 holds also when $p = \infty$. That is, if f is a function in L^∞, the standard approximating sequence $\{f_n\}$ of simple functions (see 1.5.3) converges to f uniformly, outside a set of measure 0. However, Theorem 2.4.14 fails when $p = \infty$ (see Problem 19).

If Ω is an arbitrary set, $\mathcal{S}$ a σ-algebra of subsets of Ω, and μ a complete measure, then $L^\infty(\Omega, \mathcal{S}, \mu)$ is the set of all bounded complex valued functions $f = f(x), x \in \Omega$, defined for $f \in \mathcal{S}$. The essential supremum is simply the supremum in other words, $\|f\|_\infty = \sup\{|f(x)| : x \in \Omega\}$. If Ω is the set of positive integers, $L^\infty(\Omega)$ is the space of bounded sequences of complex numbers, denoted simply by l^∞.

Problems

1. (a) If $f : [a, b] \to \mathbb{R}$, $-\infty < a < b < \infty$, is a real or complex valued function on an interval and μ is counting measure on subsets of the positive integers, show that $\int_{[a,b]} f \, d\mu = \sum_{a \le n \le b} f(n)$, where the sum is interpreted as in 2.4.12.

   (b) If $f : \Omega \to [0, \infty]$ is a real or complex valued function on the arbitrary set Ω, and μ is counting measure on subsets of Ω, show that $\int_\Omega f \, d\mu = \sum_{x \in \Omega} f(x)$, where the sum is interpreted as in 2.4.12.

2. Give an example of functions $f, f_1, f_2, \ldots$ from $\mathbb{R}$ to $[0, \infty]$ such that

   (a) each f_n is continuous on $\mathbb{R}$,

   (b) $f_n(x)$ converges to $f(x)$ for all x, $\int_{-\infty}^{\infty} f_n(x) \, dx \to \int_{-\infty}^{\infty} f(x) \, dx < \infty$, but even so $\int_{-\infty}^{\infty} |f_n(x) - f(x)| \, dx \not\to 0$, and

   (c) f is discontinuous at one point of $\mathbb{R}$.

3. (a) For each $n = 1, 2, \ldots$, let $f_n = (a_1^{(n)}, a_2^{(n)}, \ldots)$ be a sequence of complex numbers.

   (b) If the $a_k^{(n)}$ are real and $0 \le a_k^{(n)} \le a_k^{(n+1)}$ for all k and n, show that

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} a_k^{(n)} = \sum_{k=1}^{\infty} \lim_{n\to\infty} a_k^{(n)}$$

   Show that the same conclusion holds if the $a_k^{(n)}$ are complex, $\lim_{n\to\infty} a_k^{(n)}$ exists for each k, and $|a_k^{(n)}| \le b_k$ for all k and n when $\sum_{k=1}^{\infty} b_k < \infty$.

(c) If the a_j^n are real and nondecreasing, show that

$$\sum_{n=1}^{\infty}\sum_{j=1}^{\infty} a_j^n = \sum_{j=1}^{\infty}\sum_{n=1}^{\infty} a_j^n.$$

(d) If the a_j^n are complex and $\sum_{n=1}^{\infty}\sum_{j=1}^{\infty}|a_j^n| < \infty$, show that $\sum_{n=1}^{\infty}\sum_{j=1}^{\infty} a_j^n$ and $\sum_{j=1}^{\infty}\sum_{n=1}^{\infty} a_j^n$ are convergent to the same finite number.

4. Show that there is equality in the Hölder inequality if $|f|^p$ and $|g|^q$ are linearly dependent, that is, if $A|f|^p = B|g|^q$ a.e. for some constants A and B, not both 0.

5. If f is a complex-valued, nonnegative function, show that $\int_E f\,d\mu = \int_E |f|\,d\mu$ for all $E \in \mathcal{M}$ if and only if f is nonnegative a.e. on $\{x : f(x) \neq 0\}$.

6. Show that equality holds in the Cauchy–Schwarz inequality $\left|\int f\bar{g}\,d\mu\right| \leq \int |f|\,|g|\,d\mu$ if and only if f and g are linearly dependent.

7. (a) If $\mu(X) < \infty$, show that equality holds in the Minkowski inequality if and only if $Af = Bg$ a.e. for some nonnegative constants A and B, not both 0.
   (b) What are the conclusions if $\mu(X) = \infty$?

8. If $0 < r < s < t$ and $f \in L^r(\mu) \cap L^t(\mu)$, show that $f \in L^p(\mu)$ for some p between r and t. Show that $\|f\|_p \leq \max(\|f\|_r, \|f\|_t)$ for such p. Show that $\log \|f\|_p$ is a convex function of $1/p$ on $[1/t, 1/r]$. (Hint: Use Exercise 6 in Section 3.3.) Show that $\|f\|_p \to \|f\|_\infty$ as $p \to \infty$ if $\mu(X) = 1$ and $f \in L^\infty$, but this is false if $\mu(X) = \infty$.

9. If μ is finite, show that $\|f\|_r \leq \|f\|_s$ as long as $0 < r < s \leq \infty$. Give an example to show that this is false if $\mu(X) = \infty$.

10. (Jensen's inequality) If μ is a positive measure, $\mu(X) = 1$, f is a nonnegative real measurable function on X, and φ is a convex function on $(0, \infty)$, show that $\varphi\left(\int f\,d\mu\right) \leq \int \varphi \circ f\,d\mu$. (Hint: Let $s = \int f\,d\mu$. Then $\varphi(t) \geq \varphi(s) + m(t - s)$ for some m if $0 < t < \infty$.)

    Show also that the Lebesgue decomposition theorem holds if λ is a complex measure and μ is a σ-finite measure. (Use Theorem 6.10 and Section 6.13.)

11. (a) Let f be a complex-valued, μ-integrable function, where μ is a nonnegative real measure. If α is a given set of complex numbers and $|\int_E f\,d\mu| \leq \int_E |f|\,d\mu$ for all $E \in \mathcal{M}$ such that $\mu(E) > 0$, show that $\int_E f\,d\mu \in \overline{A}$ for all such E. ($\overline{A}$ is the closed disk with center at 0 and radius α.) Hint: Let $E \in \mathcal{M}$ such that $\int_E f\,d\mu \notin \overline{A}$. Choose $\gamma \in \mathbf{R}$ such that $\int_E f\,d\mu$ lies outside the disk with center $\alpha e^{i\gamma}$ and radius α. Let $H = \{z : \operatorname{Re}(z e^{i\gamma}) \leq \alpha\}$. Show that $\int_E f\,d\mu \notin H$, and conclude that $\mu(E) = 0$.

(b) If λ is a complex measure, then $\lambda \ll |\lambda|$ by definition of $|\lambda|$. Hence by the Radon–Nikodym theorem there is a $|\lambda|$-integrable function h such that $\lambda(E) = \int_E h\, d|\lambda|$ for all $E \in \mathcal{M}$. Show that $|h| = 1$ a.e. $[|\lambda|]$. [Hint: $d\nu = |h|\, d|\lambda|$ defines a finite measure ν, and we use the definition of $|\lambda|$ to show that $\lambda \geq 0$ a.e. Use part (a) to show $|h| \geq 1$ a.e.]

(c) Let μ be a nonnegative measure, possibly infinite-valued, and suppose h and $f \in L^1(\mu)$. If $A = \{x : h(x) \neq 0\}$, show that $\int_A |f| d\mu = \int_A |fh| d\mu$. (Intuitively, $\mu|_A = h\,|\mu|$ where $h = h/|h|$ and $|h| = 1$.) Deduce, as in (b), that $\int_X f(x)|h(x)| = \int_X f(x)\, dx$ if f is a bounded, complex-valued, Borel measurable function and $|h| = \overline{h}/|h|$.

(d) Under the hypotheses of (c), show that

$$\nu(E) = \int_E g\, d\mu \quad \text{for all} \quad E \in \mathcal{M}$$

13. Give an example of a bounded real-valued function f on $\mathbb{R}$ such that there is no sequence of continuous functions f_n such that $\|f - f_n\|_u \to 0$. [Hint: the continuous functions are dense in $L^{\infty}(\mathbb{R})$.]

2.5 Convergence of Sequences of Measurable Functions

In the previous section we introduced the notion of L^p convergence, and we are now familiar with convergence almost everywhere. We now consider several types of convergence and make comparisons.

Let $f, f_1, f_2, \ldots$ be complex-valued Borel measurable functions on $(\Omega, \mathcal{M}, \mu)$. We say that $f_n \to f$ in measure (or in measure μ) if for each $\varepsilon > 0$, $\mu\{x : |f_n(x) - f(x)| \geq \varepsilon\} \to 0$ as $n \to \infty$. (Notation: $f_n \xrightarrow{\mu} f$.) When μ is a probability measure, this convergence is called convergence in probability.

Our first result shows that L^p convergence is stronger than convergence in measure.

2.21 *Theorem* If $f, f_1, f_2, \ldots \in L^p(\mu)$ for $0 < p < \infty$ and $\|f_n - f\|_p \to 0$, then $f_n \xrightarrow{\mu} f$.

Proof Apply Chebyshev's inequality (2.11) to $|f_n - f|$. ∎

The same argument shows that if $\{f_n\}$ is a Cauchy sequence in L^p, then $\{f_n\}$ is Cauchy in measure; that is, for each $\varepsilon > 0$, $\mu\{x : |f_n(x) - f_m(x)| \geq \varepsilon\} \to 0$ as $n, m \to \infty$.

If $f_1, f_2, \ldots$ are complex valued Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$, we say that $f_n \to f$ almost uniformly if for given $\varepsilon > 0$ there is a set $A \in \mathcal{F}$ such that $\mu(A) < \varepsilon$ and $f_n \to f$ uniformly on A^c.

Almost uniform convergence is stronger than almost everywhere convergence and convergence in measure, as we now show.

2.5.2 *Theorem.* If $f_n \to f$ almost uniformly, then $f_n \to f$ in measure and almost everywhere.

Proof. Let $\varepsilon > 0$ be given. Since $f_n \to f$ uniformly on A^c, with $\mu(A) < \varepsilon$, if $\delta > 0$ then eventually $|f_n(x) - f(x)| < \delta$ on A^c, so $\{|f_n - f| \geq \delta\} \subset A$ and $\mu\{|f_n - f| \geq \delta\} \leq \mu(A) < \varepsilon$, proving convergence in measure.

To prove almost everywhere convergence, choose, for each positive integer k, a set A_k with $\mu(A_k) < 1/2^k$ and $f_n \to f$ uniformly on A_k^c. If $B = \bigcap_{k=1}^{\infty} A_k$, then $f_n \to f$ on B^c and $\mu(B^c) = \mu\left[\bigcup_{k=1}^{\infty} A_k\right] \leq \mu(A_k) < 1/2^k$ for all k. Thus $\mu(B^c) = 0$ and the result follows. $\square$

The converse in 2.5.2 does not hold in general, as we shall see in 2.5.4(iv), but we do have the following result.

2.5.3 *Theorem.* If $\{f_n\}$ is convergent in measure then it has a subsequence converging almost uniformly (in particular, almost everywhere) to the same limit function.

Proof. First note that $\{f_n\}$ is Cauchy in measure, because if $\{|f_n - f_m| \geq \varepsilon\} \subset \{|f_n - f| \geq \varepsilon/2\} \cup \{|f_m - f| \geq \varepsilon/2\}$, then $\mu\{|f_n - f_m| \geq \varepsilon\} \leq \mu\{|f_n - f| \geq \varepsilon/2\} + \mu\{|f_m - f| \geq \varepsilon/2\}$. Thus

$$\mu\{|f_n - f_m| \geq \varepsilon\} \leq \mu\left\{|f_n - f| \geq \frac{\varepsilon}{2}\right\} + \mu\left\{|f_m - f| \geq \frac{\varepsilon}{2}\right\} \to 0 \quad \text{as} \quad n, m \to \infty.$$

Now for each positive integer k, choose a positive integer N_k such that $N_{k+1} > N_k$ for all k and

$$\mu\{|f_n(x) - f_m(x)| \geq 2^{-k}\} < 2^{-k} \quad \text{for} \quad n, m \geq N_k.$$

Pick integers $n_k \geq N_k$ with $n_{k+1} > n_k$, $k = 1, 2, \ldots$; then if $g_k = f_{n_k}$,

$$\mu\{|g_{k+1}(x) - g_k(x)| \geq 2^{-k}\} < 2^{-k}, \quad k = 1, 2, \ldots.$$

Let $A_k = \{|g_{k+1} - g_k| \geq 2^{-k}\}$. Then $\mu(A_k) < 2^{-k}$ by the choice of n_k. Let $B = \bigcap_{k=1}^{\infty} \bigcup_{j=k}^{\infty} A_j$. Then $\mu(B) = 0$ by 2.2.4(a). Now if $x \in B^c$, then $x \in A_j^c$ eventually for any j, hence $|g_{k+1}(x) - g_k(x)| < 2^{-k}$

for $n \geq N_\varepsilon$, and it follows that $g(x) \to f(x)$. Since $g(x) = 0$ we have $\int g \, d\mu = 0$.

Thus $\int f_n \, d\mu \leq \int g \, d\mu + \varepsilon \mu(X) = \varepsilon \mu(X)$ for large n. Thus $\int f_n \, d\mu \to 0$.

(b) $\int |f_n - f| \, d\mu \leq \int |f_n - g| \, d\mu + \int |g - f| \, d\mu$. By the Weierstrass M test, $g = f$ a.e. Hence $\int |f_n - f| \, d\mu \to 0$. $\square$

Now by hypothesis, we have $f_{n_k} \to f$ for some f. Hence $f_{n_k} \to f$ a.e. By (a), $f_n \xrightarrow{\mu} g$ as well, hence $f = g$ a.e. (see Problem 1). Thus f_n converges almost uniformly to f, completing the proof. $\square$

There is a partial converse to (b) in the theorem expressing that it will be convenient to look at a condition equivalent to the convergence.

3.5.4 *Lemma.* *Suppose that $f_n \to f$ a.e. on X for every $n \in \mathbb{N}$.*

$$\mu\left\{\bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} |f_n(x) - f(x)| > \varepsilon\right\} = 0 \quad \text{for} \quad \varepsilon > 0.$$

Proof. Let $B_{n,\varepsilon} = \{x : |f_n(x) - f(x)| > \varepsilon\}$, $B_\varepsilon = \limsup_n B_{n,\varepsilon} = \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} B_{n,\varepsilon}$. Now $\bigcup_{n=1}^{\infty} B_{n,\varepsilon} \supseteq B_{n,\varepsilon}$ since $\mu(\bigcup_{n=1}^{\infty} B_{n,\varepsilon}) \geq \mu(B_{n,\varepsilon})$ so $n \to \infty$ by 3.2.7(b). Now

$$\lim_{n \to \infty} f_n(x) = f(x) \iff |f_n(x) - f(x)| < \varepsilon \text{ for large } n$$

$$\iff x \notin \bigcup_{n=1}^{\infty} B_{n,\varepsilon} \quad \text{since} \quad B_{n,\varepsilon} \supseteq B_{n+1,\varepsilon} \quad \text{for} \quad \varepsilon > 0.$$

Therefore,

$$f_n \to f \text{ a.e. iff } \mu(B_\varepsilon) = 0 \quad \text{for all} \quad \varepsilon > 0$$

$$\text{iff } \mu\left\{\bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} B_{n,\varepsilon}\right\} = 0 \quad \text{for all} \quad \varepsilon > 0. \quad \square$$

3.5.5 *Egoroff's Theorem.* *If μ is finite and $f_n \to f$ a.e. then $f_n \to f$ almost uniformly. Hence by 3.5.2, if μ is finite then almost everywhere convergence implies convergence in measure.*

Proof. It follows from 3.5.4 that given $\varepsilon > 0$ and a positive integer j, for sufficiently large $n = n(j)$, the set $B_{n(j),1/j} = \{x : |f_{n(j)}(x) - f(x)| > 1/j\}$ has measure less than $\varepsilon/2^j$. Let $A = \bigcup_{j=1}^{\infty} B_{n(j),1/j}$. Then $\mu(A) \leq \sum_{j=1}^{\infty} \mu(B_{n(j),1/j}) < \varepsilon$. Also, if $\delta > 0$

and f is chosen so that $\int_E f = 0$, we have convergence in measure and a.e. (hence a.u.), $\int_E f_n = 1$ for all n, but $\int_E |f_n - f| = 2$ for all n. Thus $\int_E f_n \not\to \int_E f$ uniformly or a.u.

We now give some examples to illustrate the relations between the various types of convergence. In all cases, we assume that E is the closure of a bounded set and μ is Lebesgue measure.

2.5.6 *Example.* (a) Let $E = [0,1]$ and define

$$f_n(x) = \begin{cases} n^2 & \text{if } 0 \le x \le \dfrac{1}{n}, \\ 0 & \text{elsewhere.} \end{cases}$$

Then $f_n \to 0$ a.e., hence in measure by 2.5.3. But for each $n \in \mathbb{N}$, f_n fails to be integrable on E. Further, as $n \to \infty$,

$$\|f_n\|_1 = \int_0^1 |f_n(x)|\,dx = n \to \infty$$

and

$$\|f_n\|_\infty = n^2 \to \infty.$$

(b) Let $E = \mathbb{R}$ and define

$$f_n(x) = \begin{cases} \dfrac{1}{n} & \text{if } n \le x \le n^2, \\ 0 & \text{elsewhere.} \end{cases}$$

Then $f_n \to 0$ uniformly on $\mathbb{R}$, so that $\int_{-\infty}^{\infty} f_n \to 0$. It follows quickly that $f_n \to 0$ a.u. and in measure. But for each $p \in [0,\infty)$, f_n fails to converge in L^p, since $\|f_n\|_p = n^{1/p} \to \infty$.

(c) Let $E = [0,1]$ and define

$$f_n(x) = \begin{cases} & \text{if } \\ 0 & \text{elsewhere.} \end{cases}$$

Then $f_n \to 0$ a.e. and in measure (as well as in L^p). One can prove, but does not immediately see, that $f_n \to 0$ uniformly on no set A with $\mu(A) > 0$. Indeed, if $f_n \to 0$ uniformly on A, then eventually $f_n = 0$ on A; hence if $A \subset [0,1]$ and $\mu(A) > 0$, then eventually $A \cap [a_n, a_n + \varepsilon_n] = \emptyset$ for sufficiently large n. Therefore, $A \cap [a_n, a_n + \varepsilon_n]$ and $\sum_{n=1}^{\infty} \mu(A \cap [a_n, a_n + \varepsilon_n]) < \infty$, a contradiction. Note that if we change

$f_n(x)$ is such that it is 1 for $n \le x \le n+1$ and 0 elsewhere, then f_n converges to 0 almost everywhere but not in measure, hence not almost uniformly.

(d): Let $\mathbb{R} = [0,1]$, and define

$$f_n(x) = \begin{cases} 1 & \text{if } \frac{n-1}{n} < x < \frac{n}{n}, \quad n = 1, \ldots, n, \quad n = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Then $\|f_n\| = 1/n \to 0$ as $n \to \infty$, so for each $\varepsilon > 0$, in the sequence $f_1, f_2, f_3, \ldots$, $f_n, f_{n+1}, f_{n+2}, \ldots$ converges to 0 in L^p. Hence converges in measure by 2.5.1. But the sequence does not converge a.e. hence by 2.5.2, does not converge almost uniformly. To see this, observe that for any $x \neq 0$, the sequence $\{f_n(x)\}$ has infinitely many zeros and infinitely many ones. Thus the set on which f_n converges has measure 0. Also, f_n does not converge in L^p, for if $f_n \xrightarrow{L^p} f$, then $f_n \xrightarrow{m} f$, hence $f = 0$ a.e. by Problem 11. But $\|f_n\|_p = 1$, a contradiction.

Problems

1. If f_n converges in both f and g in measure, show that $f = g$ a.e.

2. Show that a sequence is Cauchy in measure if it is convergent in measure.

3. (a): If μ is finite, show that L^∞ convergence implies L^p convergence for all $p \in [1, \infty)$.
   (b): Show that an extended-valued function in $M(\mathbb{R}, \mathcal{M}, \mu)$, $0 < \mu < \infty$, can be approximated in L^∞ by a polynomial, in general by a polynomial with rational coefficients.

4. If μ is finite, show that $\{f_n\}$ is Cauchy a.e. (in the sense that $\{f_n(x)\}$ is a Cauchy sequence) iff for every $\varepsilon > 0$,

$$\mu\left\{\bigcap_{n=1}^{\infty} \bigcup_{m,k \ge n} \{x : |f_m(x) - f_k(x)| \ge \varepsilon\}\right\} = 0.$$

5. Suppose g is an integrable simple function, $|f_n| \le g$ for all $n = 1, 2, \ldots$, where g is nonnegative, and $f_n \xrightarrow{\mu} f$. Show that f is integrable and $\int_E f_n \to \int_E f$.

6. A metric may be defined on the space of all measurable functions on $[0,1]$ by $\rho(f,g) = \int_0^1 \frac{|f-g|}{1+|f-g|}$ (functions that agree a.e. are identified).

(a) If $d(f_n, f) \to 0$, show that $f_n \xrightarrow{\mu} f$.

(b) If μ is finite, show that $f_n \xrightarrow{\mu} f$ implies $d(f_n, f) \to 0$.

(c) Give an example in which $f_n \xrightarrow{\mu} f$ but $d(f_n, f)$ does not approach 0.

7. If μ is a finite measure and for every $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mu\{|f_n - f| \geq \varepsilon\} < \infty$, show that $f_n \to f$ almost everywhere. Thus by Chebyshev's inequality, $\sum_{n=1}^{\infty} \|f_n - f\|_p^p < \infty$ implies that $f_n \to f$ almost everywhere.

2.6 Product Measures and Fubini's Theorem

Lebesgue measure on R^n is in a sense the product of n copies of one-dimensional Lebesgue measure, because the volume of an n-dimensional rectangle box is the product of the lengths of its sides. In this section we develop this idea in a general setting. We shall be interested in two situations. First, suppose that $(\Omega_j, \mathscr{F}_j, \mu_j)$ is a measure space for $j = 1, 2, \dots, n$. We wish to construct a measure on subsets of $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ such that the measure of the "rectangle" $A_1 \times A_2 \times \cdots \times A_n$ [with each $A_j \in \mathscr{F}_j$] is $\mu_1(A_1)\mu_2(A_2)\cdots\mu_n(A_n)$. The second situation involves compound experiments in probability. Suppose that two observations are made, with the first observation resulting in a point $\omega_1 \in \Omega_1$, the second in a point $\omega_2 \in \Omega_2$. The probability that the first observation falls into the set A is, say, $\mu_1(A)$. Furthermore, if the first observation is ω_1, the probability that the second observation falls into B is, say, $\mu(\omega_1, B)$, where $\mu(\omega_1, \cdot)$ is a probability measure defined on $\mathscr{F}_2$ for each $\omega_1 \in \Omega_1$. The probability that the first observation will belong to A and the second will belong to B should be given by

$$\mu(A \times B) = \int_A \mu(\omega_1, B)\,\mu_1(d\omega_1),$$

and we would like to construct a probability measure on subsets of $\Omega_1 \times \Omega_2$ such that $\mu(A \times B)$ is given by this formula for each $A \in \mathscr{F}_1$ and $B \in \mathscr{F}_2$. [Intuitively, the probability that the first observation will fall near ω_1 is $\mu_1(d\omega_1)$; given that the first observation is ω_1, the second observation will fall in B with probability $\mu(\omega_1, B)$. Thus $\mu(\omega_1, B)\mu_1(d\omega_1)$ represents the probability of one possible favorable outcome of the experiment. The total probability is found by adding the probabilities of favorable outcomes, in other words, by integrating over A.] Reasoning of this type may not appear natural at this point, since we have not yet talked in detail about probability theory. However, it may serve to indicate the motivation behind the theorems of this section.

3.6.1 *Definition* Let $\mathcal{T}_i$ be a σ-field of subsets of Ω_i, $i = 1, 2, \ldots, n$, and let $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. A measurable rectangle in Ω is a set $A = A_1 \times A_2 \times \cdots \times A_n$, where $A_i \in \mathcal{T}_i$ for each $i = 1, 2, \ldots, n$. The smallest σ-field containing the measurable rectangles is called the product σ-field, written $\mathcal{T}_1 \times \mathcal{T}_2 \times \cdots \times \mathcal{T}_n$. If all $\mathcal{T}_i$ coincide with a fixed σ-field $\mathcal{T}$, the product σ-field is denoted by $\mathcal{T}^n$. Note that in spite of the notation, $\mathcal{T}_1 \times \mathcal{T}_2 \times \cdots \times \mathcal{T}_n$ is not the Cartesian product of the $\mathcal{T}_i$, for Cartesian product is the set of measurable rectangles, which is the product σ-field in a certain minimal σ-field over the measurable rectangles. Notice also that the collection of finite disjoint unions of measurable rectangles forms a field (see Problem 2).

The next theorem is stated in such a way that both sets of hypotheses described above become special cases.

3.6.2 *Product Measure Theorem* Let $(\Omega_i, \mathcal{T}_i, \mu_i)$ be measure spaces with μ_i σ-finite on $\mathcal{T}_i$, and let Ω be the set $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ with σ-field $\mathcal{T}$. Assume that for each $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega$, we are given a measure $\nu(\omega_1, \omega_2, \ldots, \omega_n)(\cdot)$ on $(\Omega, \mathcal{T})$. Assume that $\nu(\omega_1, \omega_2, \ldots, \omega_n)(B)$, regarded as a function of ω for some fixed $B \in \mathcal{T}$, is a Borel measurable function on $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ for each fixed $B \in \mathcal{T}$. Assume that $\nu(\omega_1, \omega_2, \ldots, \omega_n)$ is uniformly σ-finite, that is, Ω can be written as $\bigcup_{i=1}^{\infty} B_i$, where for some positive (finite) constants c_i we have $\nu(\omega_1, \omega_2, \ldots, \omega_n)(B_i) \leq c_i$ for all $\omega \in \Omega$. (The case in which the $\nu(\omega_1, \omega_2, \ldots, \omega_n)$ are uniformly bounded, that is, $\nu(\omega_1, \omega_2, \ldots, \omega_n)(\Omega) \leq M < \infty$ for all ω, is of course included.)

Then there is a unique measure μ on $\mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \times \cdots \times \mathcal{T}_n$ such that

$$\mu(A_1 \times A_2 \times \cdots \times A_n) = \int_{A_1} \mu(\omega_1, B)(\omega_1)\, d\mu_1(\omega_1) \quad \text{for all} \quad A_1 \in \mathcal{T}_1, \quad B \in \mathcal{T}_2 \times \cdots \times \mathcal{T}_n,$$

namely,

$$\mu(F) = \int_{\Omega_1} \nu(\omega_1)(F_{\omega_1})\, d\mu_1(\omega_1) \quad F \in \mathcal{T},$$

where F_{ω_1} denotes the ω_1-section of F at ω_1:

$$F_{\omega_1} = \{(\omega_2, \ldots, \omega_n) : (\omega_1, \omega_2, \ldots, \omega_n) \in F\}.$$

Furthermore, μ is σ-finite on $\mathcal{T}$, $\nu(\omega_1)(F_{\omega_1})$ is $\mathcal{T}_1$-measurable, and all the $\nu(\omega_1)$ are probability measures on $\mathcal{T}_2 \times \cdots \times \mathcal{T}_n$.

Proof First assume that for $\nu(\omega_1, \omega_2, \ldots, \omega_n)$ we have

(*) If $C \in \mathcal{T}$, then $\nu(\omega_1, \omega_2, \ldots, \omega_n)(C) \in \mathcal{T}_1$ for each $(\omega_1, \omega_2, \ldots, \omega_n) \in \Omega$.

To prove this, let $\mathcal{C} = \{C \in \mathcal{T} : \nu(\omega_1, \omega_2, \ldots, \omega_n)(C) \in \mathcal{T}_1\}$. Then $\mathcal{C}$ is a π-field since

$$\nu\left(\omega_1, \omega_2, \ldots, \omega_n\right)\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} \nu(\omega_1, \omega_2, \ldots, \omega_n)(C_i) \text{ for } C_i \in \mathcal{C}, \quad \nu(\omega_1, \omega_2, \ldots, \omega_n)(C_i) = \nu(\omega_1)(C_i).$$

If $A \in \mathcal{M}_1$, $B \in \mathcal{M}_2$, then $(A \times B)(\omega_1) = B$ if $\omega_1 \in A$ and $\emptyset$ if $\omega_1 \notin A$. Thus $\mathcal{N}$ contains all measurable rectangles, hence $\mathcal{N} = \mathcal{M}$.

12. If $f \in \mathcal{M}$, then $g(\omega_1, \omega_2) = f(\omega_2)$ is Borel measurable on Ω.

To prove this, let $\mathcal{N}$ be the class of sets in $\mathcal{M}$ for which the conclusion of (2) holds. If $E = A \times B$, $A \in \mathcal{M}_1$, $B \in \mathcal{M}_2$, then

$$g(\omega_1, \omega_2) = \begin{cases} f(\omega_2) & \text{if } \omega_1 \in A, \\ f(\omega_2) \cdot 0 = 0 & \text{if } \omega_1 \notin A. \end{cases}$$

Thus $g(\omega_1, \omega_2) = f(\omega_2) I_A(\omega_1)$, and g is Borel measurable by hypothesis. Therefore measurable rectangles belong to $\mathcal{N}$. If $E_1, E_2, \ldots, E_n$ are disjoint measurable rectangles,

$$\mu\left[\omega_1 : \left(\bigcup_{i=1}^{n} E_i\right)(\omega_1)\right] = \mu\left[\omega_1 : \bigcup_{i=1}^{n} E_i(\omega_1)\right] = \sum_{i=1}^{n} \mu[\omega_1 : E_i(\omega_1)],$$

is a finite sum of Borel measurable functions and hence is Borel measurable in ω_2. Thus $\mathcal{N}$ contains the field of finite disjoint unions of measurable rectangles. But $\mathcal{N}$ is a monotone class, for if $E_n \in \mathcal{N}$, $n = 1, 2, \ldots$, and $E_n \uparrow E$, then $\mu(E_n)(\omega_2) \uparrow \mu(E)(\omega_2)$, by monotone convergence. Thus $\mu(E)(\omega_2) = \lim_n \mu(E_n)(\omega_2)$ is a limit of measurable functions, hence measurable in ω_2. If $E_n \downarrow E$, the same conclusion holds since $\mu(E_1)(\omega_2) < \infty$. Thus $\mathcal{N} = \mathcal{M}$.

13. Define

$$\mu(F) = \int_{\Omega_1} \mu(\omega_1, F(\omega_1))\, \mu_1(d\omega_1), \qquad F \in \mathcal{M}$$

(the integral exists by (12)). Then μ is a measure on $\mathcal{M}$, and

$$\mu(A \times B) = \int_{\Omega_1} \mu(\omega_1, B) I_A(\omega_1)\, \mu_1(d\omega_1) = \mu_1(A)\mu_2(B) \qquad \text{for all } A \in \mathcal{M}_1, \quad B \in \mathcal{M}_2.$$

To prove this, let $F_1, F_2, \ldots$ be disjoint sets in $\mathcal{M}$. Then

$$\begin{aligned} \mu\left(\bigcup_{n=1}^{\infty} F_n\right) &= \int_{\Omega_1} \mu\left[\omega_2 : \bigcup_{n=1}^{\infty} F_n(\omega_1)\right] \mu_1(d\omega_1) \\ &= \int_{\Omega_1} \sum_{n=1}^{\infty} \mu[\omega_2 : F_n(\omega_1)]\, \mu_1(d\omega_1) \\ &= \sum_{n=1}^{\infty} \int_{\Omega_1} \mu[\omega_2 : F_n(\omega_1)]\, \mu_1(d\omega_1) = \sum_{n=1}^{\infty} \mu(F_n), \end{aligned}$$

proving that μ is a measure. Now

$$\begin{aligned}\mu(A \times B) &= \int_{\Omega_1} \mu(\omega_1, A \times B)\,\nu_1(d\omega_1) \\ &= \int_{\Omega_1} \mu(\omega_1, A)\,\mu(\omega_1, B)\,\nu_1(d\omega_1) \qquad \text{(why?)} \\ &= \int_A \mu(\omega_1, B)\,\nu_1(d\omega_1),\end{aligned}$$

as desired.

Now back to the proof of Fubini's Theorem. Let $\Omega_2 = \bigcup_{n=1}^{\infty} B_n$, where the B_n's are disjoint sets in $\mathcal{F}_2$ with $\nu_2(B_n) < \infty$ for all n. If we set

$$\mu_n(\omega_1, B) = \mu(\omega_1, B \cap B_n), \qquad B \in \mathcal{F}_2,$$

then $\mu_n(\omega_1, \cdot)$ is a finite measure on $(\Omega_2, \mathcal{F}_2)$ for each $\omega_1 \in \Omega_1$, and the above construction gives a measure $\mu_n(\cdot)$ on $\mathcal{F}$ such that

$$\begin{aligned}\mu_n(A \times B) &= \int_A \mu_n(\omega_1, B)\,\nu_1(d\omega_1), \qquad A \in \mathcal{F}_1, \qquad B \in \mathcal{F}_2 \\ &= \int_A \mu(\omega_1, B \cap B_n)\,\nu_1(d\omega_1),\end{aligned}$$

namely,

$$\mu_n(F) = \int_{\Omega_1} \mu(\omega_1, F(\omega_1) \cap B_n)\,\nu_1(d\omega_1) = \int_{\Omega_1} \nu_2(F(\omega_1) \cap B_n)\,\nu_1(d\omega_1).$$

Let $\mu = \sum_{n=1}^{\infty} \mu_n$. It is left to the reader to check that μ has the desired properties.

For the uniqueness proof, assume $\mu(\omega_1, \cdot)$ is σ-finite for each ω_1. Let λ be a measure on $\mathcal{F}$ such that $\lambda(A \times B) = \int_A \mu(\omega_1, B)\,\nu_1(d\omega_1)$ for all $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$. Then $\lambda = \mu$ on the field $\mathcal{A}$ of finite disjoint unions of measurable rectangles. Now $\mu(\omega_1, \cdot)$ is σ-finite, so $\Omega_2 = \bigcup_{n=1}^{\infty} B_n$ with $B_n \in \mathcal{F}_2$ and $\mu(\omega_1, B_n) \le \infty$ for all n. Let $\Omega_1 = \bigcup_{n=1}^{\infty} A_n$, where each A_n belongs to $\mathcal{F}_1$ and $\nu_1(A_n) < \infty$. Then $\Omega_1 \times \Omega_2 = \bigcup_{n,m=1}^{\infty} (A_n \times B_m)$ and

$$\nu(A_n \times B_m) = \int_{A_n} \mu(\omega_1, B_m)\,\nu_1(d\omega_1) \le \nu_1(A_n)\,\mu(\omega_1, B_m) < \infty.$$

Thus $\lambda = \mu$ on $\mathcal{F}$ by the Caratheodory extension theorem.

We have just seen that μ is uniquely determined on $\mathcal{S}$, hence on $\mathcal{M}$. If μ and all the μ_k are probability measures, it is easy to see that μ is unique.

2.6.3 Corollary. (Finite Product Measure Theorem.) Let $(\Omega_k, \mathcal{M}_k, \mu_k)$ be a measure space for $k = 1, 2$, with μ_k σ-finite on $\mathcal{M}_k$. Let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, the set functions given by

$$\mu(P) = \int_{\Omega_1} \mu_2[P(\omega_1, \cdot)]\, d\mu_1(\omega_1) = \int_{\Omega_2} \mu_1[P(\cdot, \omega_2)]\, d\mu_2(\omega_2)$$

is the unique measure on $\mathcal{M}$ such that $\mu(A \times B) = \mu_1(A)\mu_2(B)$ for all $A \in \mathcal{M}_1$, $B \in \mathcal{M}_2$. Furthermore, μ is σ-finite on $\mathcal{M}$, and is a probability measure if μ_1 and μ_2 are. The measure μ is called the product of μ_1 and μ_2, written $\mu = \mu_1 \times \mu_2$.

Proof. ... 2.6.2, take $\mathcal{M}_k = \mu_k$ for all ω. The second formula for $\mu(P)$ is obtained by interchanging μ_1 and μ_2. ...

As a special case, Let $\Omega_1 = \Omega_2 = \mathbb{R}$, $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}(\mathbb{R})$, $\mu_1 = \mu_2 =$ Lebesgue measure. Then $\mathcal{M}_1 \times \mathcal{M}_2 = \mathcal{M}(\mathbb{R}^2)$ (Problem 2), and $\mu_1 \times \mu_2$ agrees with Lebesgue measure on intervals $[a, b] = [a_1, b_1] \times [a_2, b_2]$. By the Carathéodory extension theorem, $\mu_1 \times \mu_2$ is Lebesgue measure on $\mathcal{M}(\mathbb{R}^2)$, so we have another method of constructing two dimensional Lebesgue measure. We shall generalize to n dimensions later in the section.

The integration theory we have developed thus far includes the notion of a multiple integral on $\mathbb{R}^n$; this is simply an integral with respect to n-dimensional Lebesgue measure. However, in calculus, integrals of this type are evaluated by computing n actual integrals. The general theorem which justifies this process is Fubini's theorem, which is a direct consequence of the product measure theorem.

2.6.4 Fubini's Theorem. Assume the hypotheses of the product measure theorem 2.6.2. Let $f: (\Omega, \mathcal{M}) \to (\mathbb{R}, \mathcal{M}(\mathbb{R}))$.

(a) If $f \geq 0$, then $\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$ exists and defines a Borel measurable function of ω_1, and also

$$\int_{\Omega} f\, d\mu = \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2) \right] d\mu_1(\omega_1).$$

(b) If $\int_{\Omega} f\, d\mu$ exists in the extended sense, it follows that $\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$ exists (in the extended sense) for μ_1-almost every ω_1, and defines a Borel

measurable function of ω_2 if ν is σ-finite or if f is any third measurable function on $\Omega_1 \times \Omega_2$ that equals the integral a.e.. Also,

$$\int_{\Omega_2} g\,d\nu = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2)\,\nu(d\omega_2, d\omega_2) \right) \mu(d\omega_1, d\omega_1).$$

[The notation $\int_{\Omega_2} f(\omega_1, \omega_2)\,\nu(d\omega_2, d\omega_2)$ indicates that the function given by $g(\omega_1) = f(\omega_1, \omega_2)$ is to be integrated with respect to the measure $\nu(\omega_2, \omega_2)$.]

Theorem 5.6.1 (Fubini). Suppose that:

(1) For each fixed ω_2, the function $f(\omega_1, \omega_2) : (\Omega_1, \mathcal{A}_1) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In other words, if f is jointly measurable, then it is measurable relative to the product σ-field $\mathcal{A}_1 \times \mathcal{A}_2$ if it is measurable in each variable separately. Then if $B \in \mathcal{B}(\mathbb{R})$, $\{\omega_2 : f(\omega_1, \omega_2) \in B\} = \{\omega_2 : (\omega_1, \omega_2) \in f^{-1}(B)\} = f^{-1}(B)(\omega_1)$, where $f^{-1}(B)$ is the set in the proof of 5.6.2. Thus $\{\omega_2 : f(\omega_1, \omega_2) \in B\} \in \mathcal{A}_2$ because $f^{-1}(B) \in \mathcal{A}_1 \times \mathcal{A}_2$, as an indicator. Then

$$\int_{\Omega_2} f(\omega_1, \omega_2)\,\nu(d\omega_2, d\omega_2) = \int_{\Omega_2} f(\omega_1, \omega_2)\,\nu(d\omega_2, d\omega_2)$$

$$\mu(\omega_1)\text{-a.e. } \omega_1.$$

on. This is Borel measurable in ω_1 by part (1) of the proof of 5.6.2. Also

$$\int_{\Omega_1} f_1\,d\mu = \int_{\Omega_1} \mu(\omega_1, F(\omega_1, \omega_2))\,\nu(d\omega_2) \qquad \text{by 5.6.2}$$

$$= \int_{\Omega_1} \int_{\Omega_2} f_1(\omega_1, \omega_2)\,\mu(d\omega_1, d\omega_1)\,\nu(d\omega_2, d\omega_2).$$

Now if $f = \sum_{i=1}^{n} x_i I_{A_i}$, then by the theorem because f is a nonnegative simple function, then

$$\int_{\Omega_2} f(\omega_1, \omega_2)\,\nu(d\omega_2, d\omega_2) = \sum_{i=1}^{n} x_i \nu(\omega_1, A_i),$$

Borel measurable in ω_1

and

$$\int_{\Omega} f\,d\mu = \sum_{i=1}^{n} c_i \int_{\Omega_i} 1_{\Omega_i}\,d\mu = \sum_{i=1}^{n} c_i \int_{\Omega_1}\cdots\int_{\Omega_n} 1_{\Omega_i}(\omega_1,\ldots,\omega_n)\,d\mu_1(\omega_1)\cdots d\mu_n(\omega_n)$$

by what we have proved for indicators

$$= \int_{\Omega_1}\cdots\int_{\Omega_n} f(\omega_1,\ldots,\omega_n)\,d\mu_1(\omega_1)\cdots d\mu_n(\omega_n).$$

Finally, if $f \in \mathcal{L}^1(\Omega,\mathcal{F},\mu)$ is arbitrary, $f \geq 0$, let $f_n = f \wedge n$ be simple. Then

$$\int_{\Omega} f(\omega_1,\ldots,\omega_n)\,d\mu(\omega) = \lim_{n\to\infty} \int_{\Omega} f_n(\omega_1,\ldots,\omega_n)\,d\mu(\omega),$$

which is Borel measurable in $(\omega_1,\ldots,\omega_n)$, and

$$\int_{\Omega} f\,d\mu = \lim_{n\to\infty} \int_{\Omega} f_n\,d\mu = \lim_{n\to\infty} \int_{\Omega_1}\cdots\int_{\Omega_n} f_n(\omega_1,\ldots,\omega_n)\,d\mu_1(\omega_1)\cdots d\mu_n(\omega_n)$$

by what we have proved for simple f functions

$$= \int_{\Omega_1}\cdots\int_{\Omega_n} f(\omega_1,\ldots,\omega_n)\,d\mu_1(\omega_1)\cdots d\mu_n(\omega_n)$$

using the monotone convergence theorem twice.

This proves (a).

(b) Suppose that $\int_{\Omega} |f|\,d\mu < \infty$ as in (a).

$$\int_{\Omega_1}\int_{\Omega_2} |f(\omega_1,\omega_2)|\,d\mu_2(\omega_2)\,d\mu_1(\omega_1) = \int_{\Omega} |f|\,d\mu < \infty$$

so that $\int_{\Omega_2} |f(\omega_1,\omega_2)|\,d\mu_2(\omega_2)$ is μ_1-integrable, hence finite a.e. $[\mu_1]$, say

(2) for ω_1 almost every ω_1 we can write

$$\int_{\Omega_2} f(\omega_1,\omega_2)\,d\mu_2(\omega_2) = \int_{\Omega_2} f^+(\omega_1,\omega_2)\,d\mu_2(\omega_2) - \int_{\Omega_2} f^-(\omega_1,\omega_2)\,d\mu_2(\omega_2).$$

If $\int_{\Omega} f\,d\mu$ is finite, both integrals on the right side of (2) are finite a.e. $[\mu_1]$. In any event, we may define all integrals in (2) to be 0 for any other ω_1

measurable function of ω_1 on the exceptional set and 0 otherwise, then h will be zero for all ω_1, and will define a Borel measurable function of ω_1. If we integrate h with respect to μ_1, we obtain $\int_{\Omega_2} g(\omega_2)\,d\mu_2(\omega_2)$, and the additivity theorem for integrals

$$\int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\,d\mu_2(\omega_2)\,d\mu_1(\omega_1) = \int_{\Omega_1} h\,d\mu_1 = \int_{\Omega_1} g\,d\mu_2$$
$$= \int_{\Omega_2} g(\omega_2)\,d\mu_2(\omega_2).$$

2.6.5 *Corollary.* If $f:(\Omega_1\times\Omega_2,\mathcal{M}_1\otimes\mathcal{M}_2)\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ and the iterated integral $\int_{\Omega_1}\int_{\Omega_2}|f(\omega_1,\omega_2)|\,d\mu_2(\omega_2)\,d\mu_1(\omega_1)$ is finite, then $\int_{\Omega_1\times\Omega_2}|f|\,d\mu$ is finite and thus Fubini's theorem applies.

Proof. By 2.6.3, $\int_{\Omega_2}|f|\,d\mu_2$ is measurable, thus the measurability of 2.6.4 (ii) is satisfied. □

As a special case, we obtain the following classical result.

2.6.6 *Theorem of Fubini.* Let $(\Omega,\mathcal{M},\mu)=(\Omega_1\times\Omega_2,\mathcal{M}_1\otimes\mathcal{M}_2,\mu_1\times\mu_2)$, where μ_j is a σ-finite measure on $\mathcal{M}_j$, $j=1,2$. If f is a Borel measurable function on $(\Omega,\mathcal{M})$ such that $\int_\Omega f\,d\mu$ exists, then

$$\int_\Omega f\,d\mu = \int_{\Omega_1}\int_{\Omega_2} f\,d\mu_2\,d\mu_1$$
$$= \int_{\Omega_2}\int_{\Omega_1} f\,d\mu_1\,d\mu_2 \quad \text{by symmetry.}$$

Proof. Apply 2.6.4 with $\mu(\{\omega_1\})=\mu_1$ for all $\omega_1\in\Omega_1$. □

Note that by 2.6.5, if $\int_{\Omega_1}\int_{\Omega_2}|f|\,d\mu_2\,d\mu_1<\infty$ or $\int_{\Omega_2}\int_{\Omega_1}|f|\,d\mu_1\,d\mu_2<\infty$, the iterated integration formula 2.6.6 holds.

In 2.6.4(ii), if we wish to define $\int_{\Omega_2} f(\omega_1,\omega_2)\,d\mu_2(\omega_2)$ somewhat arbitrarily on the exceptional set where the integral does not exist, and still produce a Borel measurable function of ω_1, we could assume that $(\Omega_2,\mathcal{M}_2,\mu_2)$ is a complete measure space. The procedure is as follows. We have $(\Omega_1,\mathcal{M}_1,\mu_1)\times(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda)$, where λ is the classical Lebesgue measure on $\mathcal{M}$. We set $g(\omega_1)=h(\omega_1)$ if $g(\omega_1)=g(\omega_1)$, arbitrary $\omega_1\in A$ (by not necessarily Borel measurable). If B is a Borel subset of $\mathbb{R}$, then

$$\mu_1\{\omega_1: g(\omega_1)\in B\} = [\mu_1\{\omega_1: h(\omega_1)\in B\}\cap A^c] \cup [A\cap\{\omega_1: g(\omega_1)\in B\}]$$

The first set of the above intersection is in $\mathcal{F}_1$ and the second is a subset of A_1 which has measure zero, so it must belong to $\mathcal{F}_1$ by completeness. Thus g is Borel measurable.

In the classical Fubini theorem, we would like to define $\int_{\Omega_1} f(\omega_1, \omega_2)\,d\mu_1(\omega_1)$ and $\int_{\Omega_2} f(\omega_1, \omega_2)\,d\mu_2(\omega_2)$ as functions in a completely satisfactory fashion on the entire rectangle sets. We should assume completeness of both spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$.

The product measure theorem and Fubini's theorem may be extended to n factors as follows.

9.6.7 *Theorem.* Let $\mathcal{F}_i$ be a σ-field of subsets of Ω_i, $i = 1, \ldots, n$. Let μ_i be a σ-finite measure on $\mathcal{F}_i$, and for each $\omega_1 \in \Omega_1, \ldots, \omega_n \in \Omega_n$ let $\mu(\omega_1, \ldots, \omega_n, B)$, $B \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$, be a measure on $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Assume the $\mu(\omega_1, \ldots, \omega_n, \cdot)$ are uniformly σ-finite and assume that $\mu(\omega_1, \ldots, \omega_n, A_1 \times \cdots \times A_n)$ is measurable $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$ for each measurable rectangle $A_1 \times \cdots \times A_n$.

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$, $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

(a) There is a unique measure μ on $\mathcal{F}$ such that for each measurable rectangle $A_1 \times \cdots \times A_n \in \mathcal{F}$

$$\mu(A_1 \times \cdots \times A_n) = \int_{A_1} \mu_1(d\omega_1) \int_{A_2} \mu(\omega_1, d\omega_2) \cdots \int_{A_n} \mu(\omega_1, \ldots, \omega_{n-1}, d\omega_n) \int_{\Omega_1} \mu(\omega_1, \ldots, \omega_n, A_1 \times \cdots \times A_n)\,d\mu_1(\omega_1).$$

[Note that the last factor on the right is $\mu(\omega_1, \ldots, \omega_{n-1}, A_n)$. The assumption is on Fubini and Tonelli's probability measure $\mu(\omega_1, \ldots)$ and all the $\mu(\omega_1, \ldots, \omega_n, \cdot)$ are probability measures.]

(b) Let $f: (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $f \geq 0$ then

$$\int_\Omega f\,d\mu = \int_{\Omega_1} \mu_1(d\omega_1) \int_{\Omega_2} \mu(\omega_1, d\omega_2) \cdots \int_{\Omega_n} \mu(\omega_1, \ldots, \omega_{n-1}, d\omega_n) \int_{\Omega_1} f(\omega_1, \ldots, \omega_n)\,d\mu_1(\omega_1) \tag{11}$$

where, after the integration with respect to $\mu(\omega_1, \ldots, \omega_{k-1}, d\omega_k)$ is performed, $k = 2, \ldots, n$, the result is a Borel measurable function of $(\omega_k, \ldots, \omega_n)$.

(c) If f is quasi-integrable, or finite, then $f(\omega_1, \ldots, \omega_n)$ is for each $(\omega_1, \ldots, \omega_{n-1})$ in a set of $\mu_1 \times \cdots \times \mu_{n-1}$ measure zero, integrable with respect to $\mu(\omega_1, \ldots, \omega_{n-1}, \cdot)$ and the integral, as a function of $(\omega_1, \ldots, \omega_{n-1})$ is integrable with respect to $\mu_1 \times \cdots \times \mu_{n-1}$.

0, where λ_y is the measure determined [see (3)] by μ_y and the measures $\mu_1(y), \mu_2(y), \ldots, \mu_{n-1}(y), \ldots$. If the integral is defined on the exceptional set as 0 for any Borel measurable function on the space $\Omega_1 \times \cdots \times \Omega_n$, $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$, it becomes Borel measurable in $(y, \omega_1, \ldots, \omega_{n-1})$.

Proof. By 2.6.2 and 2.6.4, the result holds for $n = 2$. Assuming that it holds up to $n - 1$ factors, we consider the n-dimensional case. By the induction hypothesis, there is a unique measure λ_{n-1} on $\mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_{n-1}$ such that for all $A_1 \in \mathcal{F}_1, \ldots, A_{n-1} \in \mathcal{F}_{n-1}$,

$$\lambda_{n-1}(A_1 \times \cdots \times A_{n-1}) = \int_{\Omega_1} \mu_1(dy_1) \int_{A_2} \mu_2(y_1, dy_2) \cdots \int_{A_{n-1}} \mu_{n-1}(y_1, \ldots, y_{n-2}, dy_{n-1}),$$

and λ_{n-1} is σ-finite. By Exercise 9 there is a unique measure μ on $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n = (\mathcal{F}_1 \times \cdots \times \mathcal{F}_{n-1}) \times \mathcal{F}_n$ [where we write $\mathcal{F}_1 \times \cdots \times \mathcal{F}_{n-1}$ for $\mathcal{F}_1 \times (\mathcal{F}_2 \times (\cdots \times \mathcal{F}_{n-1}) \cdots)$] such that for each $A \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_{n-1}$, $B \in \mathcal{F}_n$,

$$\mu(A \times B) = \int_B \mu(y_1, \ldots, y_{n-1}, A) \lambda_n(dy_1, \ldots, dy_{n-1}) \tag{1}$$

$$= \int_{\Omega_1 \times \cdots \times \Omega_{n-1}} 1_A(y_1, \ldots, y_{n-1}) \mu(y_1, \ldots, y_{n-1}, B) \lambda_{n-1}(dy_1, \ldots, dy_{n-1}).$$

If A is a measurable rectangle $A_1 \times \cdots \times A_{n-1}$, then $\mu(y_1, \ldots, y_{n-1}, A) = 1_{A_1}(y_1) \cdots 1_{A_{n-1}}(y_{n-1})$, and (1) becomes, with the aid of the inductive hypothesis on λ_{n-1},

$$\mu(A_1 \times \cdots \times A_n) = \int_{A_1} \mu_1(dy_1) \cdots \int_{A_{n-1}} \mu_{n-1}(y_1, \ldots, y_{n-2}, dy_{n-1}) \mu(y_1, \ldots, y_{n-1}, A_n),$$

which proves the existence of the desired measure μ on $\mathcal{F}$. To show that μ is σ-finite on $\mathcal{F}$, the fact of finite additivity of measurable rectangles, and, consequently, of the unions, for $\Omega_n = \bigcup_{k=1}^{\infty} A_{nk}$, $\mu_n(A_{nk}) < \infty$, where $\mu(y_1, \ldots, y_{n-1}, A_{nk}) \le k$ for all $y_1, \ldots, y_{n-1}$, $\lambda_{n-1}(A_{nk}) < \infty$, and $\mu(A_{nk}) = \mu_n(A_{nk}) < \infty$, we have

$$\Omega = \bigcup_{k_1, \ldots, k_n = 1}^{\infty} A_{1k_1} \times \cdots \times A_{nk_n}.$$

with

$$\mu(B_1 \times B_2 \times \cdots \times B_n) = \mu_1(B_1)\mu_2(B_2)\cdots\mu_n(B_n).$$

This proves (a).

To prove (b), note that the measure μ constructed in (a) is determined by $\mu_1, \mu_2, \ldots, \mu_n$ and the measure $\mu(x_1, x_2, \ldots, x_n)$ is given by the n-fold integral

$$\int_X f\,d\mu = \int_{X_1 \times X_2 \times \cdots \times X_n} \int f(x_1, x_2, \ldots, x_n)\,d\mu(x_1, x_2, \ldots, x_n)$$

where the inner integral is Borel measurable in $x_1, x_2, \ldots, x_{n-1}$ for x_n in an exceptional set of μ_n measure 0. The desired result now follows by the induction hypothesis.

2.6.8 Theorem (a) If we take $f = \chi_E$ in formula (b) of 2.6.7(b), we obtain an explicit formula for $\mu(E)$, $E \subset X$, namely

$$\mu(E) = \int_{X_n} d\mu_n \int_{X_{n-1}} \mu(x_1, x_2, \ldots, x_{n-1}, x_n)\,d\mu_{n-1} \cdots \int_{X_1} \mu(x_1, x_2, \ldots, x_n)\,d\mu_1$$

$$\int_X f\,d\mu = \int_{X_n} d\mu_n \int_{X_{n-1}} d\mu_{n-1} \cdots \int_{X_1} f(x_1, x_2, \ldots, x_n)\,d\mu_1.$$

(b) We obtain the classical product measure and Fubini theorem by taking $\mu(x_1, x_2, \ldots, x_n) = 1$ and $\mu_j(X_j) = 1, 2, \ldots, n$. We obtain a slightly more general result that is often useful.

$$\mu(A_1 \times \cdots \times A_n) = \mu_1(A_1)\,\mu_2(A_2) \cdots \mu_n(A_n).$$

If $f \in \mathcal{L}(X, \mathcal{S}, \mu)$ and $f \geq 0$ or $\int_X f\,d\mu$ exists, then

$$\int_X f\,d\mu = \int_{X_n} d\mu_n \int_{X_{n-1}} d\mu_{n-1} \cdots \int_{X_1} f\,d\mu_1$$

and the iterations may be performed in any order. The measure μ is called the product of $\mu_1, \ldots, \mu_n$ and we write $\mu = \mu_1 \times \cdots \times \mu_n$. In particular, if each μ_j is Lebesgue measure on $\mathcal{S}(R^1)$, then $\mu_1 \times \cdots \times \mu_n$ is Lebesgue measure on $\mathcal{S}(R^n)$, just as in the discussion after 2.6.1.

Problems

1. Show that the collection of all disjoint unions of measurable rectangles in $X_1 \times \cdots \times X_n$ forms a field

2 Show that $\mathscr{B}(\mathbb{R}) \times \cdots \times \mathscr{B}(\mathbb{R})$ is finite.

3 If $\mathscr{F}_1, \ldots, \mathscr{F}_n$ are algebras of subsets, show that

$$(\mathscr{F}_1 \times \cdots \times \mathscr{F}_{n-1}) \times \mathscr{F}_n = \mathscr{F}_1 \times (\mathscr{F}_2 \times \cdots \times \mathscr{F}_n).$$

4 Let μ be the product of the σ-finite measures μ_1 and μ_2 in $(\Omega_1, \mathscr{F}_1, \Omega_2, \mathscr{F}_2)$. Show that the following are equivalent.

(a) $\mu(A) = 0$.
(b) $\mu_2(\{\omega_2 : \omega_1 \in A\}) = 0$ for μ_1-almost all $\omega_1 \in \Omega_1$.
(c) $\mu_1(\{\omega_1 : \omega_2 \in A\}) = 0$ for μ_2-almost all $\omega_2 \in \Omega_2$.

5 In Problem 4, let $(\Omega_1, \mathscr{F}_1, \mu_1)$ be the completion of $(\Omega_1, \mathscr{F}_1, \mu_1)$ and assume μ_2 is complete. If $A \in \mathscr{F}_1$, show that $\omega_1 \in \Omega_1$ for μ_1-almost all $\omega_1 \in \Omega_1$, $\{\omega_2 : (\omega_1, \omega_2) \in A\} \in \mathscr{F}_2$ for μ_2-almost all $\omega_2 \in \Omega_2$. Give an example in which $A \in \mathscr{F}_1$ but $\{\omega_2 : (\omega_1, \omega_2) \in A\} \notin \mathscr{F}_2$ for some $\omega_1 \in \Omega_1$.

6 (a) Let $\Omega_1 = \Omega_2$ = the set of positive integers, $\mathscr{F}_1 = \mathscr{F}_2 =$ all subsets, $\mu_1 = \mu_2$ = counting measure. $f(\omega_1, \omega_1) = \omega_1$, $f(\omega_1, \omega_1 + 1) = -\omega_1$, $\omega_1 = 1, 2, \ldots$, $f(\omega_1, \omega_2) = 0$ if $\omega_2 \neq \omega_1$ or $\omega_1 + 1$. Show that $\int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2 = 0$, $\int_{\Omega_1} \int_{\Omega_2} f\, d\mu_2\, d\mu_1 = +\infty$. (Tonelli's theorem fails since $\int_{\Omega_1} \int_{\Omega_2} |f|\, d\mu_2\, d\mu_1 = +\infty$.)

(b) Let $\Omega_1 = \Omega_2 = [0, 1]$, $\mathscr{F}_1 = \mathscr{F}_2 = \mathscr{B}([0, 1])$, $\mu_1 =$ Lebesgue measure, $\mu_2 =$ counting measure. Let $A = \{(\omega_1, \omega_2) : \omega_1 = \omega_2\} \in \mathscr{F}_1 \times \mathscr{F}_2$. Show that

$$\int_0^1 \int_0^1 I_A\, d\mu_2\, d\mu_1 = \int_0^1 \mu_2(\{\omega_1\})\, d\mu_1(\omega_1) = +\infty$$

but

$$\int_0^1 \int_0^1 I_A\, d\mu_1\, d\mu_2 = \int_0^1 \mu_1(\{\omega_2\})\, d\mu_2(\omega_2) = 0.$$

(Fubini's theorem fails since μ_1 is not σ-finite; the proof is incorrect. The theorem fails also since $\int_{\Omega_1} |\mu_2|$ is not $\mathscr{F}_1$-measurable.)

$$\int_0^1 \mu_1(\{\omega_2\})\, d\mu_2(\omega_2) \text{ is not } \mathscr{F}_1\text{-measurable.}$$

7 Let $\Omega_1 = \Omega_2$ = the first uncountable ordinal, $\mathscr{F}_1 = \mathscr{F}_2$ = all subsets, $\Omega = \Omega_1 \times \Omega_2$, $\mathscr{F} = \mathscr{F}_1 \times \mathscr{F}_2$. Assume the continuum hypothesis which identifies Ω_1 and Ω_2 with $[0, 1]$.

(a) If f is any function from Ω^n to $\mathbb{R}$ for a subset $A \subseteq \Omega$ to $[0,1]$ and $G = \{(x, y) : x \in \Omega, y = f(x)\}$ is the graph of f, show that $f \in \mathcal{F}^n$.

(b) Let $\Omega = \{(x, y) \in \Omega^2 : y < x\}$, $\mathcal{F} = \{(x, y) \in \Omega^2 : y \le x\}$. For $x \in \Omega$ let $\mathcal{B} = \{y : (x, y) \in \Omega\}$. Show that $\mathcal{B} \in \mathcal{F}$. (The relation $y \le x$ refers to the ordering of $\mathbb{R}$ and is not known, not real numbers.

(c) Show that $\mathcal{F}$ contains all subsets of Ω.

8 Show that a measurable function of measurable is jointly measurable. Specifically, if $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ are given, $f : \Omega_1 \to \Omega_2$, and $\mathcal{F}^1$ be $f(\omega_1, \omega_2) = g(\omega_1)$, then f is measurable relative to $\mathcal{F}_1 \times \mathcal{F}_2$ and $\mathcal{F}^1$, regardless of the nature of $\mathcal{F}_2$.

9 Give an example of a function $f : [0,1] \times [0,1] \to [0,1]$ such that

(a) $f(x, y)$ is Borel measurable in x for each fixed y and Borel measurable in y for each fixed x,

(b) f is not jointly measurable; that is, f is not measurable relative to the product σ-field $\mathcal{B}[0,1] \times \mathcal{B}[0,1]$, and

(c) $\int_0^1 \left[\int_0^1 f(x, y)\,dy\right]dx$ and $\int_0^1 \left[\int_0^1 f(x, y)\,dx\right]dy$ exist but are unequal. (This example is suggested by Problem 7.)

2.7 Measurability in Infinite Product Spaces

The measurability problem in stochastic processes arises in the following manner. Suppose we have a single probability space $(\Omega, \mathcal{F}, P)$ which is adequate to describe the outcome of a single experiment, but we are interested in the outcomes of an infinite sequence of such experiments. Can we construct a probability space which is adequate to describe the outcomes of the infinite sequence of experiments? Such a space can be constructed if the product measure theorem can be extended to infinitely many dimensions. Our task is to extend the product of finitely many σ-fields, but we must consider the infinitely many infinite case.

2.7.1 *Definition.* For each $i = 1, 2, \ldots$ let $(\Omega_i, \mathcal{F}_i)$ be a measurable space. Let $\Omega = \prod_{i=1}^{\infty} \Omega_i$ be the set of all sequences $(\omega_1, \omega_2, \ldots)$ such that $\omega_i \in \Omega_i$, $i = 1, 2, \ldots$. For $n \ge 1$ and $\prod_{i=1}^{n} \Omega_i$, we define

$$\mathcal{B}_n = \{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B^n\}.$$

The set $\mathcal{B}_n$ is called the cylinder with base B^n, the cylinder is said to be measurable if $B^n \in \prod_{i=1}^{n} \mathcal{F}_i$. If $B^n = A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{F}_i$, for each i, $\mathcal{B}_n$ is called a measurable *rectangle*. A measurable rectangle if $A_i \in \mathcal{F}_i$ for each i.

A cylinder with an n-dimensional base may always be regarded as having a higher dimensional base. For example, if

$$B^n = \{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B^n\}$$

that

$$B = \{\omega \in \Omega : (\omega_{i_1}, \ldots, \omega_{i_n}) \in C^1, \omega \in \Omega\}$$
$$= \{\omega \in \Omega : (\omega_{j_1}, \ldots, \omega_{j_n}) \in C^2, \omega \in \Omega\}.$$

It follows that the measurable cylinders form a field. It is also the case that finite unions of measurable rectangles form a field. The argument is the same as in Problem 1 of Section 2.6.

The minimal σ-field over the measurable cylinders is called the product of the σ-fields $\mathcal{F}_i$, written $\prod_{i=1}^{\infty} \mathcal{F}_i$ or $\prod_{i=1}^{\infty} \mathcal{F}_i$. It is also the minimal σ-field over the measurable rectangles (see Problem 11). If all $\mathcal{F}_i$ coincide with a fixed σ-field $\mathcal{F}_1$, then $\prod_{i=1}^{\infty} \mathcal{F}_i$ is denoted by $\mathcal{F}_1^{\infty}$, and in all Ω_i coincide with a fixed set S, $\prod_{i=1}^{\infty} \Omega_i$ is denoted by S^{∞}.

The infinite-dimensional version of the product measure theorem will be given only for probability measures and for coordinate spaces of real sequences. (In fact the construction to be described below runs into trouble for σ-probability measures.)

24.2 Theorem. Let $(S_1, \mathcal{F}_1, P_1), (S_2, \mathcal{F}_2, P_2), \ldots$ be probability measure spaces. Let $\Omega = \prod_{i=1}^{\infty} S_i$, $\mathcal{F} = \prod_{i=1}^{\infty} \mathcal{F}_i$. Suppose that we are given a consistent sequence of finite-dimensional probability measures P_n on $\mathcal{F}_n$, and for each $j = 1, 2, \ldots$ and each $(\omega_1, \ldots, \omega_j) \in S_1 \times \cdots \times S_j$ we are given a probability measure $P(\omega_1, \ldots, \omega_j, \cdot)$ on $\mathcal{F}_{j+1}$. Assume that $P(\omega_1, \ldots, \omega_j, C)$ is measurable $\mathcal{F}_1 \times \cdots \times \mathcal{F}_j$ as a function of $(\omega_1, \ldots, \omega_j)$ for each fixed $C \in \mathcal{F}_{j+1}$.

If $B \in \prod_{i=1}^{n} \mathcal{F}_i$, define

$$P_{n+1}(B) = \int_{S_1} P_1(d\omega_1) \int_{S_2} P(\omega_1, d\omega_2) \cdots \int_{S_n} P(\omega_1, \ldots, \omega_{n-1}, d\omega_n)$$
$$\int_{S_n} I_B(\omega_1, \ldots, \omega_n) P(\omega_1, \ldots, \omega_n, d\omega_{n+1}).$$

Note that P_{n+1} is probability measure on $\prod_{i=1}^{n+1} \mathcal{F}_i$ by 24.1 and 24.3(a).

There is a unique probability measure P on $\mathcal{F}$ such that for all n, P agrees with P_n on $\prod_{i=1}^{n} \mathcal{F}_i$; that is, $P\{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B\} = P_n(B)$ for all $n = 1, 2, \ldots$ and all $B \in \prod_{i=1}^{n} \mathcal{F}_i$.

Proof. Any measurable cylinder can be represented in the form $B_n = \{\omega \in \Omega : (\omega_1, \ldots, \omega_n) \in B^n\}$ for some n and some $B^n \in \prod_{i=1}^{n} \mathcal{F}_i$. Define $P(B_n) = P_n(B^n)$. We must show that P is well defined on measurable cylinders. For suppose that B_n can also be represented as $\{\omega \in \Omega : (\omega_1, \ldots, \omega_m) \in C^m\}$ where $C^m \in \prod_{i=1}^{m} \mathcal{F}_i$. We must show that $P_n(B^n) = P_m(C^m)$. Say

$n > m$. Then $(\omega_1, \ldots, \omega_n) \in \Omega^n$ iff $(\omega_1, \ldots, \omega_m) \in \Omega^m$, hence $\Omega^n = \Omega^m \times \Omega_{m+1} \times \cdots \times \Omega_n$. It follows from the definition of P_n that $P_n(B \times \Omega_{m+1} \times \cdots \times \Omega_n) = P_m(B)$ if $B \subset \Omega^m$. (The fact that the $P(\omega_1, \ldots, \omega_n)$ are probability measures is used here.)

Since P_n is a measure on $\prod_1^n \mathcal{A}_i$, it is immediate that P is finitely additive on the field $\mathcal{F}_0$ of measurable cylinders. To show that P is continuous from above at the empty set, it suffices (by Problem 2.19) to prove that P is countably additive on $\mathcal{F}_0$, and the Carathéodory extension theorem extends P to a probability measure on $\prod_1^\infty \mathcal{A}_i$. By construction, P agrees with P_n on n-dimensional cylinders.

Let $\{B_n, n = 1, 2, \ldots\}$ be a sequence of measurable cylinders decreasing to $\emptyset$. For any $\omega \in \Omega_1$ and $n \geq 1$, let $B_n^\omega = \{(\omega_2, \ldots, \omega_n) : (\omega, \omega_2, \ldots, \omega_n) \in B_n\}$. Take $B_1 = \Omega$ for all n. Assume $\lim_{n \to \infty} P(B_n) > 0$. Then for each $n \geq 1$,

$$P(B_n) = \int_{\Omega_1} p_1^{(n)}(\omega_1) p(\omega_1, d\omega_2),$$

where

$$p_1^{(n)}(\omega_1) = \int_{\Omega_2} P(\omega_2, d\omega_3) \cdots \int_{\Omega_n} p(\omega_{n-1}, d\omega_n) I_{B_n^{\omega_1}}(\omega_2, \ldots, \omega_n).$$

Since $B_{n+1} \subset B_n$, it follows that $B_{n+1}^{\omega_1} \subset B_n^{\omega_1} \times \Omega_{n+1}$, hence

$$I_{B_{n+1}^{\omega_1}}(\omega_2, \ldots, \omega_{n+1}) \leq I_{B_n^{\omega_1}}(\omega_2, \ldots, \omega_n).$$

Therefore $p_1^{(n)}(\omega_1)$ decreases as n increases to $I_{B_1^{\omega_1}}(\omega_2, \ldots)$, say $p_1^{(n)}(\omega_1) \downarrow p_1(\omega_1)$. By the bounded monotone convergence theorem, $P(B_n) = \int_{\Omega_1} p_1^{(n)}(\omega_1) P(d\omega_1) \downarrow \int_{\Omega_1} p_1(\omega_1) P(d\omega_1)$. If $\lim_{n \to \infty} P(B_n) > 0$, then $p_1(\omega_1) > 0$ for some $\omega_1 \in \Omega_1$. In fact, $\omega_1 \in \bar{B}_1$, for if $\omega_1 \notin \bar{B}_1$, then $p_1^{(n)}(\omega_1) = 0$ for all n, hence $p_1(\omega_1) = 0$ for all n, and $p_1(\omega_1) = 0$, a contradiction.

Now for each $n \geq 2$,

$$p_2^{(n)}(\omega_2) = \int_{\Omega_2} p(\omega_2, d\omega_3) \cdots \int_{\Omega_n} p(\omega_{n-1}, d\omega_n) I_{B_n^{\omega_1, \omega_2}}(\omega_3, \ldots, \omega_n),$$

where

$$p_2^{(n)}(\omega_2) = \int_{\Omega_3} P(\omega_3, d\omega_4) \cdots \int_{\Omega_n} p(\omega_{n-1}, d\omega_n) I_{B_n^{\omega_1, \omega_2}}(\omega_3, \ldots, \omega_n)$$

As above, $\varphi_J^I(\omega_J) = \omega_I$ if $\omega_J \in \Omega_J$ and $\omega_I \in \Omega_I$ for $I \subset J$ finite. Hence

$$\varphi_J^I(\omega_J) = \int_{\Omega_K} \omega_K(d\omega_K) \quad \text{for } K = J \setminus I.$$

Since $\varphi_J^I(\omega_J) = \omega_I(\varphi_I^I(\omega_J))$ for $\omega_J \in \Omega_J$ and $\omega_I \in \Omega_I$ and $\omega_I(\Omega_I) = 1$ for $\omega_I \in \Omega_I$, we have $\omega_I(B_I) = 1$ for some $B_I \in \mathcal{B}_I$ and $\omega_J(\varphi_J^I(B_I)) = 1$ for $\omega_J \in B_J$.

The process may be repeated inductively to obtain points $\omega_{J_1}, \omega_{J_2}, \ldots$ such that for each n, $\omega_{J_n}(\varphi_{J_n}^{J_{n-1}}(\cdots(\varphi_{J_2}^{J_1}(B_{J_1}))\cdots)) = 1$ for some $B_{J_1} \in \mathcal{B}_{J_1}$. This is a contradiction. This proves the existence of the desired probability measure P. If Q is another such probability measure, then $P = Q$ on measurable cylinders, hence $P = Q$ on $\mathcal{F}$ by the uniqueness part of the Carathéodory extension theorem. □

The classical product measure theorem extends as follows.

2.7.3 Corollary. *For each $j = 1, 2, \ldots$, let $(\Omega_j, \mathcal{F}_j, P_j)$ be an arbitrary probability space. Let $\Omega = \prod_{j=1}^{\infty} \Omega_j$, $\mathcal{F} = \prod_{j=1}^{\infty} \mathcal{F}_j$. Then there is a unique probability measure P on $\mathcal{F}$ such that*

$$P(\{\omega \in \Omega : \omega_{j_1} \in A_{j_1}, \ldots, \omega_{j_n} \in A_{j_n}\}) = \prod_{i=1}^{n} P_{j_i}(A_{j_i})$$

for all $n = 1, 2, \ldots$ and all $A_{j_i} \in \mathcal{F}_{j_i}$, $i = 1, 2, \ldots, n$. We call P the product of the P_j, and write $P = \prod_{j=1}^{\infty} P_j$.

Proof. In 2.7.2 take $P(\omega_{j_1}, \ldots, \omega_{j_n}) = P_{j_1} \times \cdots \times P_{j_n}$. Then $P(\{\omega : \omega_{j_1} \times \cdots \times \omega_{j_n}\}) = \prod_{i=1}^{n} P_{j_i}(A_{j_i})$, and thus the probability measure P of 2.7.2 has the desired properties. If Q is another such probability measure, then $P = Q$ on the field of finite disjoint unions of measurable rectangles, hence $P = Q$ on $\mathcal{F}$ by the Carathéodory extension theorem. □

Thus far we have considered probability measures on infinitely many product spaces. The results may be extended to uncountable products if certain assumptions are made about the individual factor spaces. We will be completely general at the beginning, but when we reach the main result, the Kolmogorov extension theorem, we will assume that all the factor spaces are the reals with the Borel field of Borel sets.

2.7.4 Definitions and Comments. *Suppose T is an arbitrary index set. For each $t \in T$, let $(\Omega_t, \mathcal{F}_t)$ be a measurable space. Let $\prod_{t \in T} \Omega_t$ be the set of all functions $\omega = \{\omega(t) : t \in T\}$ such that $\omega(t) \in \Omega_t$ for each $t \in T$. If $t_1, \ldots, t_n \in T$,*

$\mathcal{B}^n \subset \prod_{t\in T} \Omega_t$, we define the set $B(t_1, \ldots, t_n)$ as $\{\omega \in \prod_{t\in T} \Omega_t : (\omega(t_1), \ldots, \omega(t_n)) \in B\}$. We call $B(t_1, \ldots, t_n)$ the cylinder with base B on $(t_1, \ldots, t_n)$. The cylinder is said to be measurable if $B^n \subset \prod_{t\in T} \mathcal{B}_t$. If $B = B_1 \times \cdots \times B_n$ the cylinder is called a *rectangle*. A measurable rectangle is T if $B_t = \Omega_t$, $t \neq t_1, \ldots, t_n$. Note that if $\Omega_t = \Omega$ then $\prod_{t\in T} \Omega_t = \Omega^T$, the set of all functions from T to Ω.

For example, let $T = [0, 1]$, $\Omega = \mathbb{R}$, for all $t \in T$, $\mathcal{B} = \{(a, b) : -\infty < a < b < \infty\}$. Then

$$B = \left\{\omega : \omega\left(\tfrac{1}{2}\right) \in \left(-2, 3\right), \omega\left(\tfrac{1}{4}\right) > 0, \omega\left(\tfrac{3}{4}\right) < 2\right\}$$

see Fig. 2.7.1, where $A_1 = (-2, 3)$ and $A_2 = (-\infty, 0)$, $A_3 = (-\infty, 2)$.

Exactly as in 2.7.1, the measurable cylinders form a field, as do the finite disjoint unions of measurable rectangles. The minimal σ-field over the measurable cylinders is denoted by $\prod_{t\in T} \mathcal{B}_t$ and called the product of the σ-fields $\mathcal{B}_t$. If $\Omega_t = \Omega$ and $\mathcal{B}_t = \mathcal{B}$ for all t, $\prod_{t\in T} \mathcal{B}_t$ is denoted by $\mathcal{B}^T$, again as in 2.7.1. $\prod_{t\in T} \mathcal{B}_t$ is also the minimal σ-field over the measurable rectangles.

We now consider the problem of constructing probability measures on $\prod_{t\in T} \mathcal{B}_t$. The approach will be as follows: Let $\tau = \{t_1, \ldots, t_n\}$ be a finite subset of T, where $t_1 < t_2 < \cdots < t_n$. Let T in such a subset of $\mathbb{R}$, we are given a probability measure P_τ on $\prod_{i=1}^n \mathcal{B}_{t_i}$; $P_\tau(B)$ is to represent $P[(\omega(t_1), \ldots, \omega(t_n)) \in B]$. We shall require that the P_τ be "consistent"; to see what kind of consistency is needed, consider an example.

Suppose T is the set of positive integers and $\Omega_t = \mathcal{B}_t = \mathcal{B}(\mathbb{R})$ for all t; suppose we know $P_{\{1,2,3,4\}}(B^4) = P[(\omega_1, \omega_2, \omega_3, \omega_4) \in B^4]$ for

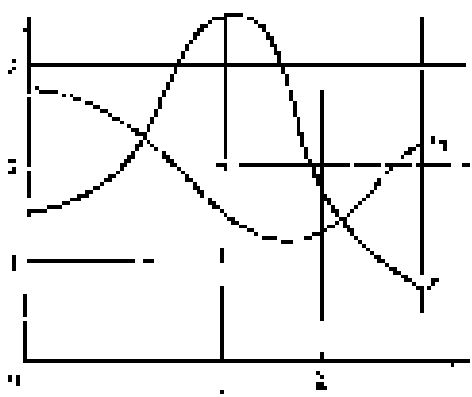

Figure 2.7.1.

all $B^k \in \mathcal{B}(\mathbb{R}^k)$. Then $P\{\omega : (\omega_{t_1}, \ldots, \omega_{t_k}) \in B^k\} = P_{t_1,\ldots,t_k}(B^k \times \mathbb{R} \times \mathbb{R}^1 \times \mathbb{R}^1 \cdots) = \mu_{t_1,\ldots,t_k}(\mathbb{R}^k \times B^k \times \mathbb{R}^1 \cdots) = \mu_{t_1,\ldots,t_k}(B^k)$, $B^k \in \mathcal{B}(\mathbb{R}^k)$. Thus more probabilities of sets involving just the first k coordinates are essential, probabilities of sets involving ω_{t_1} and ω_{t_2} as well as ω_{t_1} and so on are also determined. Thus the original space has lots of P's that agree with the measures induced from $P_{t_1,\ldots,t_k}$. We hope to find a consistent family of probability measures $P_{t_1,\ldots,t_k}$ which determines a unique probability measure on $\prod_{t\in T} \mathbb{R}_t$.

Now to formalize. Let $\mathbb{T} = \{(t_1, \ldots, t_k) : t_1 < \cdots < t_k, \text{ for some } k \in \mathbb{N}\}$. For $\mathbb{S} \subset T$, $\prod_{t\in\mathbb{S}} \mathbb{R}_t$ is denoted by $\mathbb{R}_{\mathbb{S}}$. If $\mathbb{S} = \{t_1, \ldots, t_k\}$ is a nonempty subset of $\mathbb{T}$ and $\mathbb{S} = \{(t_1, t_2, \ldots, t_k) \in \mathbb{T}$, then the projection $\pi_{\mathbb{S}}(\omega)$ of $\omega \in \mathbb{R}_{\mathbb{S}}$ is denoted by $\omega_{\mathbb{S}}$. Similarly if $\omega = (\omega_t)_{t\in T} \in \prod_{t\in T} \mathbb{R}_t$ then $\omega_{\mathbb{S}}$ for $\mathbb{S} \subset T$ will be used for $(\omega_{t_1}, \ldots, \omega_{t_k})$. If $\mathbb{S} = T$, the measurable cylinder sets based $\mathcal{B}$ will be written as $\mathcal{B}(\mathbb{R})$.

If P is a probability measure on $\mathcal{B}_T$, the projection of P on $\mathcal{B}_{\mathbb{S}}$ is the probability measure $\pi_{\mathbb{S}}(P)$ on $\mathcal{B}_{\mathbb{S}}$ defined by

$$[\pi_{\mathbb{S}}(P)](B) = P\{\omega \in \mathbb{R}_T : \omega_{\mathbb{S}} \in B\} \qquad B \in \mathcal{B}_{\mathbb{S}}.$$

Similarly, if Q is a probability measure on $\prod_{t\in\mathbb{S}} \mathbb{R}_t$, the projection of Q on $\mathcal{B}_{\mathbb{S}}$ is defined by

$$[\pi_{\mathbb{S}}(Q)](B) = Q\left\{\omega \in \prod_{t\in\mathbb{S}} \mathbb{R}_t : \omega_{\mathbb{S}} \in B\right\} = Q(B \times \prod_{t\in\mathbb{S}-\mathbb{S}} \mathbb{R}_t), \qquad B \in \mathcal{B}_{\mathbb{S}}.$$

Our main result, the Kolmogorov extension theorem, can be proved when each $\mathbb{R}_t$ is a complete, separable metric space, with $\mathcal{B}_t$ the class of Borel sets (the smallest σ-field generated by the open sets). However, to avoid certain technical complications, we will take all $\mathbb{R}_t$ to be $\mathbb{R}$ and $\mathcal{B}_t = \mathcal{B}(\mathbb{R})$.

2.7.5 *Kolmogorov Extension Theorem.* For each n in the arbitrary index set T, let $\mathbb{R}_t = \mathbb{R}$ and $\mathcal{B}_t$ the Borel sets on $\mathbb{R}$.

Assume that for each finite nonempty subset $\mathbb{S}$ of T we are given a probability measure $P_{\mathbb{S}}$ on $\mathcal{B}_{\mathbb{S}}$. Assume that the $P_{\mathbb{S}}$ are consistent, that is, $\pi_{\mathbb{S}}(P_{\mathbb{T}}) = P_{\mathbb{S}}$ for each nonempty $\mathbb{S} \subset \mathbb{T}$.

Then there is a unique probability measure P on $\mathcal{B} = \prod_{t\in T} \mathcal{B}_t$ such that $\pi_{\mathbb{S}}(P) = P_{\mathbb{S}}$ for all $\mathbb{S} \subset \mathbb{T}$.

Proof. We define the desired measure on measurable cylinders by $P(B^{\mathbb{S}} \times \mathbb{R}_{T-\mathbb{S}}) = P_{\mathbb{S}}(B^{\mathbb{S}})$, $B^{\mathbb{S}} \in \mathcal{B}_{\mathbb{S}}$.

We must show that this definition makes sense since a given measurable cylinder can be represented in several ways. For example, suppose that

$B^n = [a_n, b_n) \times \cdots$. Then

$$\begin{aligned} P^n(B^n) &= \mu_n([a_n, b_n)) \cdots \mu_1([a_1, b_1)) \\ &= (b_n - a_n) \cdots (b_1 - a_1) \\ &= P^n(B^n \times \mathbb{R} \times \mathbb{R} \times \cdots) \quad \text{where} \quad B^n = [a_n, b_n) \times [a_{n-1}, b_{n-1}) \times \cdots \end{aligned}$$

It is sufficient to consider dual representations of the same measurable cylinder in two forms $B^n(\omega) = B^k(\omega)$ where $k < n$ and $n < \infty$. But then

$$\begin{aligned} P_n(B^n) &= [\mu_n P_{n-1}](B^{n-1}) \quad \text{by the consistency hypothesis} \\ &= P_{n-1}(B^{n-1}) \quad \text{by definition of projection} \end{aligned}$$

But this assumption $B^n(\omega) = B^k(\omega)$ simply means that if $x \in \Omega_n$, then $x \in B^n$ iff $y_n = \pi^k_n(x) \in B^k$, hence $P_n(B^n) = P_{n-1}(B^{n-1})$ as desired.

Thus P is well-defined on measurable cylinders. The class $\mathcal{C}$ of measurable cylinders clearly is a field, and $\sigma(\mathcal{C}) = \mathcal{F}$.

Now if $A_1, \ldots, A_m$ are disjoint sets in $\mathcal{F}_n$, we may write (by introducing extra factors as in the above example) $A_i = B^n_i(\omega)$, $i = 1, \ldots, m$, where $B^n_i = [a_{i1}, b_{i1}) \times \cdots \times [a_{in}, b_{in}) \times \mathbb{R} \times \mathbb{R} \times \cdots$ are disjoint sets in $\mathcal{F}_n$. Thus

$$\begin{aligned} P\left[\bigcup_{i=1}^m A_i\right] &= P\left[\bigcup_{i=1}^m B^n_i(\omega)\right] \\ &= P_n\left[\bigcup_{i=1}^m B^n_i\right] \quad \text{by definition of } P \\ &= \sum_{i=1}^m P_n(B^n_i) \quad \text{since } P_n \text{ is a measure and} \\ &= \sum_{i=1}^m P(A_i) \quad \text{again by definition of } P. \end{aligned}$$

Therefore P is finitely additive on $\mathcal{F}_n$. To show that P is countably additive on $\mathcal{F}_n$, we must verify that P is continuous from above at $\emptyset$ and invoke Theorem 1.3.3(b). The Carathéodory extension theorem (1.3.6) then extends P to $\mathcal{F}$.

Let $A_j = B^n_j(\omega)$ be a sequence of measurable cylinders decreasing to $\emptyset$. If $P(A_j)$ does not approach 0, we have, for some $\varepsilon > 0$, $P(A_j) \geq \varepsilon$ for all j. Suppose $A_j = B^{n_j}_j(\omega)$; by taking an extra factor $\mathbb{R}$ if necessary we may assume that the numbers n_j are strictly increasing with j.

It follows from (1.4.1) that we can find a compact set $C_j \subset B^{n_j}_j$ such that $P_{n_j}(B^{n_j}_j - C_j) < \varepsilon/2^{j+1}$. Define $A'_j = C_j(\omega) \cap A_j$. Then $P(A_j - A'_j)$

$= P_{x_0}(B^{n-1} \times C^{n-1} \times \{x_n\}) \cdot 2\varepsilon 2^{-n}$. In this way we approximate the given cylinder by cylinders with compact bases.

Now note

$$B_n = A_1^{(n)} \times A_2^{(n)} \times \cdots \times A_n^{(n)} \subseteq A_1^{(n)} \times A_2^{(n)} \times \cdots \times A_{n-1}^{(n)} = B_{n-1},$$

so

$$P_{x_0}(B_n \setminus B_{n-1}) = P\left[\left(\bigcap_{i=1}^{n} A_i^{(n)}\right) \setminus \left(\bigcap_{i=1}^{n-1} A_i^{(n-1)}\right)\right] \leq \sum_{i=1}^{n} P(A_i^{(n)} \setminus A_i^{(n-1)})$$

$$+ \sum_{i=1}^{n-1} P(A_i^{(n)} \setminus A_i^{(n-1)}) < \sum_{i=1}^{n} \varepsilon 2^{-i} + \varepsilon 2^{-n} < \varepsilon.$$

Since $B_n \setminus A_n^{(n)} \subseteq A_n^{(n)} \setminus A_n$, $P(A_n^{(n)} \setminus B_n) = P(B_n \setminus A_n^{(n)}) < \varepsilon 2^{-n}$, consequently $P(B_n \setminus A_n) < \varepsilon 2^{-n+1}$. In particular, for B_1 is nonempty.

Now, pick $x_1 \in B_1$, $x_2 \in A_2^{(2)}, \ldots$. Let $A_1^{(1)} = [0, x_1]$, $A_2^{(1)} = (x_1, 1]$; note that $B_1 = A_1^{(1)}$. Consider the sequence

$$\left(A_1^{(1)}, \ldots, A_n^{(1)}\right), \left(A_1^{(2)}, \ldots, A_n^{(2)}\right), \left(A_1^{(3)}, \ldots, A_n^{(3)}\right), \ldots$$

that is, $\gamma_1, \gamma_2, \gamma_3, \ldots$.

Since the γ_i belong to C^n, a compact subset of Ω_n, we have a convergent subsequence approaching some $\gamma_0 \in C^n$. Let $\gamma_0 = (A_1^{(0)}, \ldots, A_n^{(0)})$ for $A_i^{(0)} \subseteq A_i$ for $i \geq 2$. Consider the sequence $\gamma_0^{(1)}, \gamma_0^{(2)}, \ldots \in C^n$ (eventually, one exists a convergent subsequence $\gamma_0^{(i)} \to \gamma_0 \in C^n$).

Note that $\gamma_0^{(i)} = A_1^{(i)}$ since $x_1 \in B_1$. Thus the left-side approaches $[0, x_1]$, so $A_1^{(0)} = [0, x_1]$ necessarily as the right side approaches x_1. Hence $A_1^{(0)} = A_1^{(1)}$.

Continuing in this fashion, at stage n we have a subsequence

$$\gamma_{n-1}^{(i)} \to \gamma_{n-1} = (A_1^{(n-1)}, \ldots, A_{n-1}^{(n-1)}) \quad \text{and} \quad (A_1^{(n-1)}, \ldots, A_{n-1}^{(n-1)}) = \gamma_{n-1} \quad \text{for} \quad i \geq 1.$$

Picking $\omega \in \prod_{i=1}^{\infty} \Omega_i$ such that $\omega_i = x_i$ for $i = 1, 2, \ldots$ such a choice is possible since $\omega_1 = x_1 \in A_1^{(1)}$. Then $\omega_n \in A_n^{(n)}$ for each n hence

$$\omega \in \bigcap_{n=1}^{\infty} A_n^{(n)} \subseteq \bigcap_{n=1}^{\infty} A_n = A,$$

a contradiction. Thus P extends to a measure on $\mathcal{F}$ and by construction $m_{x_0}(B) = P_{x_0}(B)$ for all B.

Finally, if P and Q are two probability measures on $\mathcal{F}$ such that $\pi_t(P) = \pi_t(Q)$ for all finite $t \subset T$, then for any $\beta \in \mathcal{T}_0$,

$$P(\beta^{-1}(A)) = \lim_{t \downarrow \beta} P(\pi_t^{-1}(\pi_t(\beta^{-1}(A)))) = \lim_{t \downarrow \beta} Q(\pi_t^{-1}(\pi_t(\beta^{-1}(A)))) = Q(\beta^{-1}(A)).$$

Thus P and Q agree on measurable cylinders, and hence on $\mathcal{F}$ by the unique part of the Caratheodory extension Theorem 1.

Problems

1. Show that $\prod_{t \in T} \mathcal{B}_t$ is the minimal σ-field over the measurable rectangles.
2. Let $\mathcal{F} = \mathcal{B}(\mathbb{R}^T)$, show that the following sets belong to $\mathcal{F}$:
   (a) $\{x \in \mathbb{R}^T : \sup_{t \in T} x_t \le a\}$,
   (b) $\{x \in \mathbb{R}^T : \lim_{t \in T, t \downarrow \beta} x_t \le a\}$,
   (c) $\{x \in \mathbb{R}^T : \lim_{t \in T, t \downarrow \beta} x_t \text{ exists and is finite}\}$,
   (d) $\{x \in \mathbb{R}^T : \limsup_{t \in T, t \downarrow \beta} x_t \le a\}$, and
   (e) $\{x \in \mathbb{R}^T : \sum_{t \in T} x_t^2 < \infty$ (Cauchy criterion$)\}$.
3. Let $\mathcal{F}$ be a collection of subsets of a set Ω, and assume $\mathcal{F}$ is countably generated, that is, there is a sequence $\{A_1, A_2, \ldots\}$ in $\mathcal{F}$ such that the smallest σ-field containing all the A_i is $\mathcal{F}$. Show that $\mathcal{F}^{[0,1]}$ is countably generated. In particular, $\mathcal{B}(\mathbb{R})^{[0,1]}$ is countably generated, where the A_i are intervals with rational endpoints.
4. Prove that any set in $\mathcal{B}(\mathbb{R})^{[0,1]}$ is countably determined.
5. Let $f : \mathbb{R}^{[0,1]} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \text{the smallest positive integer } n \\ \text{such that } x_{\frac{1}{n}} + \cdots + x_{\frac{n}{n}} \ge 1, \\ \text{if such an } n \text{ exists,} \\ \infty \text{ if } x_{\frac{1}{n}} + \cdots + x_{\frac{n}{n}} < 1 \text{ for all } n. \end{cases}$$

   Show that $f : (\mathbb{R}^{[0,1]}, \mathcal{B}(\mathbb{R})^{[0,1]}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

2.8 Weak Convergence of Measures

In 2.5 we studied convergence of sequences of measurable functions. We now examine a somewhat different notion of convergence. The results form the starting point for the study of the central limit theorem of probability.

We will need some properties of semicontinuous functions; proofs of all necessary results are given in Appendix 2.

If Ω is a metric space, the class of Borel sets of Ω, denoted by $\mathcal{B}(\Omega)$, is defined to be the σ-field generated by the open subsets of Ω. In our applications to probability, we will need only the case $\Omega = \mathbb{R}^1$.

3.A.7 **Theorem.** Let $\mu, \mu_1, \mu_2, \ldots$ be Borel measures on the Borel sets of a metric space Ω. The following conditions are equivalent.

(a) $\int_\Omega f\,d\mu_n \to \int_\Omega f\,d\mu$ for all bounded continuous $f : \Omega \to \mathbb{R}$.

(b) $\liminf_{n\to\infty} \int_\Omega f\,d\mu_n \geq \int_\Omega f\,d\mu$ for all bounded lower semicontinuous $f : \Omega \to \mathbb{R}$.

(c) $\limsup_{n\to\infty} \int_\Omega f\,d\mu_n \leq \int_\Omega f\,d\mu$ for all bounded upper semicontinuous $f : \Omega \to \mathbb{R}$.

(d) $\int_\Omega f\,d\mu_n \to \int_\Omega f\,d\mu$ for all bounded $f : \Omega \to \mathbb{R}$ such that $\mu(\{x \in \Omega : f \text{ is discontinuous at } x\}) = 0$ [a].

(e) $\liminf_{n\to\infty} \mu_n(A) \geq \mu(A)$ for every open set $A \subset \Omega$, and $\mu_n(\Omega) \to \mu(\Omega)$.

(f) $\limsup_{n\to\infty} \mu_n(A) \leq \mu(A)$ for every closed set $A \subset \Omega$ and $\mu_n(\Omega) \to \mu(\Omega)$.

(g) $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{F}(\Omega)$ such that $\mu(\partial A) = 0$ [b] (where ∂A denotes the boundary of A).

Proof (a) $\Rightarrow$ (b): If $a \in \mathbb{R}$ and f is bounded continuous,

$$\liminf_{n\to\infty} \int_\Omega f\,d\mu_n \geq \liminf_{n\to\infty} \int_\Omega g\,d\mu_n = \int_\Omega g\,d\mu \qquad \text{by (a)}$$

But since f is lower semicontinuous, (3.8.7), it is the limit of an sequence of continuous functions, and if $f > -M$, we may take the g's to be bounded below by $-M$. In any case, $\int g\,d\mu$ is an increasing sequence whose limit may be $+\infty$. Thus if we take the sup over g in the above equation, we obtain (b).

(b) $\Leftrightarrow$ (c): Note that f is LSC $\Leftrightarrow$ $-f$ is upper semicontinuous $\Leftrightarrow$ $-f \in \mathrm{USC}(\Omega)$.

(b) $\Rightarrow$ (d): Let $\underline{f}$ be the lower semicontinuous envelope of f, the sup of all LSC functions g such that $g \leq f$. Since $f \geq 0$, $\underline{f}(x) = \liminf_{y\to x} f(y)$ and $f(x) = \limsup_{y\to x} f(y)$. Consequently if f is discontinuous at x then $\underline{f}(x) < f(x) = \overline{f}(x)$. Furthermore, $\underline{f}$ is LSC and $\overline{f}$ is USC. Thus if f is bounded and continuous a.e. $[\mu]$,

$$\begin{aligned}
\int_\Omega f\,d\mu = \int_\Omega \underline{f}\,d\mu &\leq \liminf_{n\to\infty} \int_\Omega \underline{f}\,d\mu_n \qquad \text{by (b)} \\
&\leq \liminf_{n\to\infty} \int_\Omega f\,d\mu_n \qquad \text{since } \underline{f} \leq f \\
&\leq \limsup_{n\to\infty} \int_\Omega f\,d\mu_n \leq \limsup_{n\to\infty} \int_\Omega \overline{f}\,d\mu_n \text{ since } f \leq \overline{f} \\
&\leq \int_\Omega \overline{f}\,d\mu \qquad \text{by (b)} \\
&= \int_\Omega f\,d\mu \qquad \text{(proving (d))}
\end{aligned}$$

(c) $\Rightarrow$ (a): Clearly (c) implies (a), which in turn implies (b). If (b) is true, then T_2 is $\mathcal{A}(C)$, so by (b), $\limsup_{n\to\infty} \mu(A \cap T_2^{-n}B) \geq \mu(A)\mu(B)$. Now $T_2 = T_1$ on $\mathcal{A}(\mathcal{S}) = \mathcal{A}(\mathcal{S})$ by (c).

(c) $\Rightarrow$ (c): Take complements.

(c) $\Rightarrow$ (a): Replace A by $T^k A$ in the statement of (a), k the integer of (c). Then

$$\limsup_{n\to\infty} \mu(A \cap T^{-n}B) = \limsup_{n\to\infty} \mu(T^k A \cap T^{-n}B) = \mu(A)\mu(B) \qquad \text{by (c)}$$

$$= \mu(A)\mu(B) \qquad \text{by hypothesis.}$$

Also, using (a),

$$\liminf_{n\to\infty} \mu(A \cap T^{-n}B) \geq \liminf_{n\to\infty} \mu(A \cap T^{-n}B) = \mu(A)\mu(B) = \mu(A)\mu(B)$$

(a) $\Rightarrow$ (c): Let f be a bounded continuous function on Ω. If $|f| < M$, let $A_i = \{x \in X : \mu(f^{-1}[i,i+1)) \neq 0\}$. A_i is countable since the $f^{-1}[i,i+1)$ are disjoint and μ is finite. Construct a partition of $[-M,M]$, say $-M = y_0 < y_1 < \cdots < y_n = M$ where $y_i \notin A_i$ ($i = 0, 1, \ldots, n$) (M may be increased if necessary). Let $B_i = \{x : y_i \leq f(x) < y_{i+1}\}$, $i = 0, 1, \ldots, n-1$. It follows from (a) that

$$\sum_{i,j=0}^{n-1} y_i y_j \mu(B_i \cap T^{-k}B_j) \to \sum_{i,j=0}^{n-1} y_i y_j \mu(B_i)\mu(B_j)$$

as $k \to \infty$, $k \in \mathcal{S}$ (or equivalently $k \in [0,\infty)$). Now

$$\int_\Omega f(x)\,dx = \int_\Omega f(x)\,d\mu = \left| \int_\Omega f(x)\,dx - \sum_{i,j=0}^{n-1} y_i \mu(B_i) \right|$$

$$+ \left| \sum_{i,j=0}^{n-1} y_i \mu(B_i) - \sum_{i,j=0}^{n-1} y_i y_j \mu(B_i \cap T^{-k}B_j) \right|$$

$$+ \left| \sum_{i,j=0}^{n-1} y_i y_j \mu(B_i \cap T^{-k}B_j) - \int_\Omega f(x)\,dx \right|$$

The first and third terms on the right may be written as

$$\left| \sum_{i=0}^{n-1} \int_{B_i} (f(x) - y_i)\,d\mu \right| \leq \mu(\Omega)\,\varepsilon,$$

and this is bounded by $\max_{1\le i\le n} \mu(\{x_i\})$, which can be made arbitrarily small by choice of the partition since $\mu(\{x\}) = 0$ and $\mu(\Omega) < \infty$ can be [illegible]. The third term on the right is bounded by $\max_{1\le i\le n} \mu(\{x_i\})$, which can also be made arbitrarily small. The second term approaches 0 as $n \to \infty$, proving [illegible]. ∎

2.8.2 *Theorem.* Another condition equivalent to those of 2.8.1 is: that $\int_\Omega f\,d\mu_n \to \int_\Omega f\,d\mu$ for all bounded uniformly continuous $f : \Omega \to \mathbb{R}$ [illegible].

The convergence described in 2.8.1 is sometimes called weak convergence of measures. We shall write $\mu_n \to \mu$.

If the measures μ_n and μ are defined on $\mathcal{B}(\mathbb{R})$, then we have corresponding distribution functions F_n and F on $\mathbb{R}$. We may relate convergence of measures to convergence of distribution functions.

2.8.3 *Definition.* A continuity point of a distribution function F on $\mathbb{R}$ is a point $x \in \mathbb{R}$ such that F is continuous at x. [illegible] (thus by convention, ∞ and $-\infty$ are discontinuity points).

2.8.4 *Theorem.* Let $\mu, \mu_1, \mu_2, \ldots$ be finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with corresponding distribution functions $F, F_1, F_2, \ldots$. The following are equivalent:

(a) $\mu_n \to \mu$;

(b) $F_n(x) \to F(x)$ at all continuity points x of F, where $F(x) = \mu((-\infty, x])$, $F(-\infty) = \lim_{x\to-\infty} F(x)$, $F(+\infty) = \lim_{x\to\infty} F(x)$.

If all distribution functions are 0 at $-\infty$, condition (b) is equivalent to the statement that $F_n(x) \to F(x)$ for all points $x \in \mathbb{R}$ at which F is continuous, and $F_n(x) \to F(x)$.

Proof. (a) $\Rightarrow$ (b). If a and b are continuity points of F in $\mathbb{R}$, then $\partial[a,b]$ is a Borel set whose boundary has μ measure 0. By 2.8.1 (b), $\mu_n([a,b]) \to \mu([a,b])$, that is, $F_n(b) - F_n(a) \to F(b) - F(a)$. If $a = -\infty$, the argument is the same, and if $\mu_n(\{\infty\}) = \mu(\{\infty\}) = 0$ then ∞ is a point of the boundary of a Borel set whose boundary has μ measure 0, and the proof proceeds as before.

(b) $\Rightarrow$ (a). Let A be an open subset of $\mathbb{R}$; write it as the disjoint union of open intervals $I_1, I_2, \ldots$. Then

$$\liminf_{n\to\infty} \mu_n(G) = \liminf_{n\to\infty} \sum_k \mu_n(\{x_k\})$$

$$\geq \sum_k \liminf_{n\to\infty} \mu_n(\{x_k\}) \qquad \text{by Fatou's lemma.}$$

Let $\varepsilon > 0$ be given. For each k, let N_k be a finite semiclosed subinterval of A_k such that the endpoints of N_k are continuity points of F and $\mu(N_k) \geq \mu(A_k) - \varepsilon 2^{-k}$. (If A_k is a semiinfinite interval, N_k is necessarily semifinite by the choice of ε.) Then

$$\liminf_{n\to\infty} \mu_n(G) \geq \liminf_{n\to\infty} \mu_n(\bigcup_k N_k) = \mu(\bigcup_k N_k) \qquad \text{for all } k.$$

Thus

$$\liminf_{n\to\infty} \mu_n(G) \geq \sum_k \mu(N_k) \geq \sum_k \mu(A_k) - \varepsilon = \mu(G) - \varepsilon.$$

Since ε is arbitrary, we have $\mu_n \xrightarrow{v} \mu$ by 2.8.1(ii). □

Condition (b) of 2.8.4 is sometimes called weak convergence of the sequence $\{\mu_n\}$ or $\{F_n\}$, written $F_n \xrightarrow{w} F$.

Problems

1 (a) If S is a closed subset of the metric space Ω, show that $\mu_n \xrightarrow{v} \mu$ if and only if $\int_S f\,d\mu_n \to \int_S f\,d\mu$ for all bounded uniformly continuous $f: \Omega \to \mathbb{R}$.

(b) Show that in 2.8.1, $\mu_n \xrightarrow{v} \mu$ if and only if $\int_\Omega f\,d\mu_n \to \int_\Omega f\,d\mu$ for all bounded uniformly continuous $f: \Omega \to \mathbb{R}$.

2 Show that in 2.8.1, $\mu_n \xrightarrow{v} \mu$ if and only if $\mu_n(A) \to \mu(A)$ for all open sets A such that $\mu(\partial A) = 0$.

2.9 Bibliography

The presentation in Chapters 1 and 2 has been strongly influenced by several sources. The first systematic presentation of measure theory appeared in Halmos (1950). Halmos achieves slightly greater generality by the construction of outer measures by exhausting a field by sets of finite measure, (A sequence of sets $\{A_n\}$ is said to exhaust a set A if $A_n \subset A_{n+1}$ for all n and $\bigcup_n A_n = A$.) Halmos' treatment of outer measures and countable additivity is also elegant. However, the fields will be unnecessarily adequate for our purposes. The first account of measure theory specifically oriented toward probability was given by Doob (1953).

Several useful refinements were made by Bergman (1961), Nirenberg (1959), and Rudin (1956). Nirenberg's book emphasizes pseudoconvexity, while Rudin's book is primarily devoted to a preparation for work in harmonic analysis.

The Cartan theorems on finitely additive set functions and the development of integral representations for functions with values in a Banach space, see Dunford and Schwartz (1958).

A more recent treatment of currents and cohomology theory is that of Hörmander (1973), who discusses applications to Fourier analysis and probability. A development of measure theory that emphasizes its connection with probability is given by Doob (1994).

CHAPTER

3

INTRODUCTION TO FUNCTIONAL ANALYSIS

3.1 Banach spaces

An important part of analysis concerns the study of vector spaces endowed with additional structure of some kind. In Chapter 2, for example, we studied the vector space $L^1(\Omega, \mathcal{F}, \mu)$. If $1 \le p \le \infty$, the seminorm $\|\cdot\|_p$, defined on $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$, gives rise to the normed vector spaces $L^p(\Omega, \mathcal{F}, \mu)$. In this chapter, we will study various structures that can be defined on vector spaces. The most general structure which we will not study in detail is that of a topological vector space, which is a vector space endowed with a topology in which the vector space operations, that is, the mappings $(x, y) \mapsto x + y$ and $(\alpha, x) \mapsto \alpha x$, are continuous. We will concern ourselves with two special types, Banach and Hilbert spaces. In a Banach space, there is a notion of length of a vector, and in a Hilbert space, length is in turn determined by a "dot product" of vectors. If X and Y are Banach spaces, then $X \times Y$ is a natural generalization of the finite-dimensional Euclidean spaces.

We now begin the topics we are going to study. The term "vector space" will always mean vector space over the complex field $\mathbb{C}$. "real vector space" indicates that the scalar field is $\mathbb{R}$. Other fields will be considered.

3.1.1 Definitions. Let X be a vector space. A seminorm on X is a function $\|\cdot\| : X \to [0, \infty)$ which satisfies

$$\|\alpha x\| = |\alpha| \, \|x\| \qquad \text{for all } \alpha \in \mathbb{C},\ x \in X,$$

$$\|x + y\| \le \|x\| + \|y\| \qquad \text{for all } x, y \in X.$$

The first property is called absolute homogeneity, the second subadditivity. Note that absolute homogeneity implies that $\|0\| = 0$. We use the same symbol for the zero vector and the zero scalar; if $\alpha \in \mathbb{C}$, $|\alpha|$ is simply the value of α. The seminorm is called a norm on X if $\|x\| > 0$ whenever $x \neq 0$. Normed spaces

If $\|\ \|$ is a seminorm on L and $d(x,y) = \|x - y\|$, $x, y \in L$, then d is a pseudometric on L, a metric if $\|\ \|$ is a norm on L. A Banach space is a complete normed linear space, that is, relative to the metric d induced by the norm every Cauchy sequence converges.

An inner product on L is a function $(\ ,\ ) : L \times L \to \mathbb{C}$, denoted by $(x, y) \mapsto (x, y)$, satisfying

$$(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z) \quad \text{for } \alpha, \beta \in \mathbb{C},\ x, y, z \in L,$$

$$(x, y) = \overline{(y, x)} \quad \text{for } x, y \in L,$$

$$(x, x) \geq 0 \quad \text{for all } x \in L,$$

$$(x, x) = 0 \quad \text{if and only if } x = 0.$$

(the over bar indicates complex conjugation). A vector space endowed with an inner product is called an inner product space or pre-Hilbert space. If L is an inner product space, $\|x\| = (x, x)^{1/2}$ defines a norm on L; this is a consequence of the Cauchy–Schwarz inequality, to be proved in Section 3.1 below, with this norm L is called a Hilbert space. Thus a Hilbert space is a Banach space whose norm is determined by an inner product.

Finally, a topological vector space is a vector space L with a topology such that the addition and scalar multiplication are continuous, in other words, the mappings

$$(x, y) \mapsto x + y \quad \text{of } L \times L \text{ into } L$$

and

$$(\alpha, x) \mapsto \alpha x \quad \text{of } \mathbb{C} \times L \text{ into } L$$

are continuous, with the product topology on $L \times L$ and $\mathbb{C} \times L$.

A Banach space is a topological vector space with the topology induced by the metric $d(x, y) = \|x - y\|$, but if $x_n \to x$ and $y_n \to y$ then

$$\|x_n + y_n - (x + y)\| \leq \|x_n - x\| + \|y_n - y\| \to 0,$$

if $x_n \to x$ and $\alpha_n \to \alpha$ then

$$\begin{aligned}\|\alpha_n x_n - \alpha x\| &\leq |\alpha_n - \alpha|\,\|x_n - x\| + |\alpha|\,\|x_n - x\| \\ &= |\alpha_n - \alpha|\,\|x_n - x\| + |\alpha_n - \alpha|\,\|x\| \to 0.\end{aligned}$$

The above definitions remain unchanged if L is a real vector space, except of course that $\mathbb{C}$ is replaced by $\mathbb{R}$. Also we may drop the conjugate in the symmetry requirement for inner products and simply write $(x, y) = (y, x)$ for $x, y \in L$.

3.1.2 Examples. (a) $L^p(\Omega, \mathcal{F}, \mu)$ is a normed vector space and $\|\cdot\|_p$ is a seminorm on the vector space $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$. If we pass to equivalence classes by identifying functions that agree a.e., we obtain $L^p(\Omega, \mathcal{F}, \mu)$, a normed space (see 3.1). When $p = 2$, the norm $\|\cdot\|_2$ is determined by an inner product:

$$\langle f, g \rangle = \int_\Omega f g \, d\mu.$$

Thus $L^2(\Omega, \mathcal{F}, \mu)$ is a Hilbert space.

(b) $\mathcal{F}$ consists of all subsets of Ω and μ is counting measure, that is, $\mu(A) = n$ if A has n elements, $\mu(A) = \infty$ if A is infinite. $L^p(\Omega, \mathcal{F}, \mu)$ is a Banach space denoted $l^p(\Omega)$ or $l^p(\Gamma)$ if $\Gamma = \Omega$.

By 3.1.1(b), if $1 \le p < \infty$, then $l^p(\Omega)$ consists of all functions $f = \{f(\omega) : \omega \in \Omega\}$ from Ω to $\mathbb{C}$ such that $f(\omega) = 0$ for all but countably many ω, and $\|f\|_p = \left[\sum_\omega |f(\omega)|^p\right]^{1/p} < \infty$. When $p = 2$ the norm on $l^2(\Omega)$ is induced by the inner product

$$\langle f, g \rangle = \sum_\omega f(\omega) \overline{g(\omega)}.$$

When $p = \infty$ the situation is slightly different. The space $l^\infty(\Omega)$ is the collection of all bounded complex-valued functions on Ω with the sup norm

$$\|f\| = \sup\{|f(\omega)| : \omega \in \Omega\}.$$

Similarly, $l^\infty\Omega$ is a normed space for more generally, a topological space and C the space of all bounded continuous complex-valued functions on Ω, then C is a Banach space under the sup norm. For we may verify directly that the sup norm is actually a norm, or equivalently, we may note that $C \subseteq l^\infty(\Omega)$. This we needed only check completeness, and this follows because a uniform limit of continuous functions is continuous.

(c) Let c be the set of all convergent sequences of complex numbers, and put the sup norm on it. If $f = \{a_n : n \ge 1\} \in c$, we then

$$\|f\| = \sup\{|a_n| : n = 1, 2, \ldots\}.$$

Again, to show that c is a Banach space we need only establish completeness. Let $\{f_k\}$ be a Cauchy sequence in c, so if $f_k = \{a_{k,n} : n \ge 1\} \in c$, then for each $\varepsilon > 0$ there exists N such that $\|f_k - f_l\| < \varepsilon$ if $k, l \ge N$. Thus $|a_{k,n} - a_{l,n}| < \varepsilon$ for all n, where $a_{k,n} = a_n$ and $b_l = \lim_{n \to \infty} a_{l,n}$ exists uniformly in k, since $|a_{k,n} - a_{l,n}| < \|f_k - f_l\| < \varepsilon$ as $n, m \to \infty$. By the standard double limit theorem,

$$\lim_{k \to \infty} \lim_{n \to \infty} a_{k,n} = \lim_{n \to \infty} \lim_{k \to \infty} a_{k,n}.$$

In particular, $\lim_{n\to\infty} x_n = a$ exists, so let $y = \lim_{n\to\infty} x_n \in [a,b]$. Then $f(x_n) \to f(y)$ since f is continuous, but $|f(x_n) - f(y)| = \sup_{x\in[a,b]} |x - y| = 0$ is impossible since $x_n \to y$. This proves completeness.

(c) (Could you find a topological proof? You may skip this exercise.)

Let L be the collection of all complex-valued functions on X whose support is countable. Put the discrete topology on $\mathbb{C}$, so that a sequence in L of functions in L converges to the function $f \in L$ if and only if $\lim_{n\to\infty} f_n(x) = f(x)$ for each $x \in X$. With this topology, L is a topological vector space. To show that addition is continuous, observe that if $f_n \to f$ and $g_n \to g$ pointwise, then $f_n + g_n \to f + g$ pointwise. Similarly $cf_n \to cf$, $n = 1, 2, \ldots$, $c_n \to c$, and $f_n \cdot g_n \to f \cdot g$ pointwise, thanks to $f_n \to f$ pointwise.

3.3 Real Banach Spaces and Hilbert Spaces

We begin our study of normed vector spaces with a discussion of finite-dimensional Euclidean spaces, as the sense that many of the familiar geometric results in $\mathbb{R}^n$ carry over. First recall the definition of the inner product for "dot product" on $\mathbb{R}^n$. If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, then $\langle x, y\rangle = \sum_{j=1}^n x_j y_j$. (The notation $\sum_{j=1}^n x_j y_j$ in the space ℓ^2 of all sequences of complex numbers, the length of a vector in $\mathbb{R}^n$ is given by $\|x\| = (\sum_{j=1}^n x_j^2)^{1/2} = (\sum_{j=1}^n |x_j|^2)^{1/2}$ and the distance between two points of $\mathbb{R}^n$ is $d(x,y) = \|x - y\|$. In order to show that it is a metric, the triangle inequality must be established. This in turn follows from the Cauchy–Schwarz inequality $|\langle x, y\rangle| \le \|x\|\,\|y\|$. In fact, the Cauchy–Schwarz inequality holds in any inner product space, as we now prove:

3.3.1 Cauchy–Schwarz Inequality. *If L is an inner product space, and $\|x\| = \langle x, x\rangle^{1/2}$, $x \in L$, then*

$$|\langle x, y\rangle| \le \|x\|\,\|y\| \qquad \text{for all} \qquad x, y \in L.$$

Equality holds if x and y are linearly dependent.

Proof. For any $x, y \in L$,

$$0 \le \|x - \alpha\langle x, y\rangle y\|^2 = \langle x - \alpha\langle x, y\rangle y,\ x - \alpha\langle x, y\rangle y\rangle$$
$$= \|x\|^2 + \alpha^2|\langle x, y\rangle|^2\|y\|^2 - 2\alpha|\langle x, y\rangle|^2.$$

Since $\alpha = -|\langle x, y\rangle|^2/\|y\|^2 \in \mathbb{R}$, then $x = 0$ and the result is trivial. Since $\langle x, y\rangle = \langle y, x\rangle$, we have

$$0 \le \|x - \alpha\langle x, y\rangle y\| = \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2} + \frac{|\langle x, y\rangle|^2}{\|y\|^2},$$

proving the inequality.

As $|x+ty|^2 = \|x\|^2 + 2t\,\mathrm{Re}\langle x,y\rangle + t^2\|y\|^2$ is ≥ 0 for all t, the discriminant $4(\mathrm{Re}\langle x,y\rangle)^2 - 4\|x\|^2\|y\|^2 \leq 0$, and the first inequality follows. If equality holds and $x \neq 0$, then $|x+ty|^2$ has a real root t, so $x+ty=0$ and x and y are linearly dependent. ∎

3.2.2 *Corollary.* If V is an inner product space and $\|x\| = \langle x,x\rangle^{1/2}$, $x \in V$, then $\|\ \|$ is a norm on V.

Proof. It is immediate that $\|x\| \geq 0$, $\|ax\| = |a|\,\|x\|$, and $\|x\| = 0 \Leftrightarrow x = 0$. Now

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \|x\|^2 + \langle x,y\rangle + \langle y,x\rangle + \|y\|^2 \\ &= \|x\|^2 + 2\,\mathrm{Re}\langle x,y\rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2|\langle x,y\rangle| + \|y\|^2 \qquad \text{by 3.2.1} \\ &\leq \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.\end{aligned}$$

Therefore $\|x+y\| \leq \|x\| + \|y\|$. ∎

3.2.3 *Corollary.* An inner product is (jointly) continuous in both variables, that is, $x_n \to x$, $y_n \to y$ implies $\langle x_n, y_n\rangle \to \langle x,y\rangle$.

Proof.

$$\begin{aligned}|\langle x_n,y_n\rangle - \langle x,y\rangle| &= |\langle x_n - x, y_n\rangle + \langle x, y_n - y\rangle| \\ &\leq \|x_n - x\|\,\|y_n\| + \|x\|\,\|y_n - y\| \qquad \text{by 3.2.1}\end{aligned}$$

Therefore, by the boundedness of the norm, $\|x_n\| \leq K$ for some K and all n, so $|\langle x_n,y_n\rangle - \langle x,y\rangle| \leq K\|x_n - x\| + \|x\|\,\|y_n - y\|$, and therefore,

$$\|x_n\| - \|x\| \leq \|x_n - x\| + \|x\| - \|x\|,$$

hence $\|x_n\| - \|x\| \leq \|x_n - x\|$. It follows that $\langle x_n, y_n\rangle \to \langle x,y\rangle$. ∎

The computation of 3.2.2 establishes the following result, which says intuitively that the sum of the squares of the lengths of the diagonals of a parallelogram is twice the sum of the squares of the lengths of the sides.

3.2.4 *Parallelogram law.* In an inner product space

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

Proof.

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\langle x,y\rangle,$$

and

$$\|x - v\|^2 = \|x\|^2 + \|v\|^2 - 2\,\mathrm{Re}\langle x, v\rangle.$$

Now suppose that $v_1, \ldots, v_n$ are mutually perpendicular unit vectors in $\mathbb{C}^n$ (or $\mathbb{R}^n$). If x is an arbitrary vector in $\mathbb{C}^n$, we try to approximate x by a linear combination $\sum_{k=1}^{n} \alpha_k v_k$. The reader may recall that $\sum_{k=1}^{n} \alpha_k v_k$ will be closest to x in the sense of Euclidean distance when $\alpha_k = \langle x, v_k\rangle$. This result holds in an arbitrary inner product space:

5.2.5 Definition Two elements x and y of an inner product space V are said to be orthogonal or perpendicular if $\langle x, y\rangle = 0$. If $U \subseteq V$, $x \in V$, and x is orthogonal to every $y \in U$, then x is said to be orthogonal to U. A set $\{x_\alpha : \alpha \in A\}$ of vectors in V is orthogonal if $x_\alpha \neq 0$ for all $\alpha \in A$ and $\langle x_\alpha, x_\beta\rangle = 0$ whenever $\alpha \neq \beta$. It is orthonormal if it is orthogonal and $\|x_\alpha\| = 1$ for all $\alpha \in A$.

An argument similar to 5.2.3 shows that if $x_1, \ldots, x_n$ are orthogonal, we have the Pythagorean relation $\left\|\sum_{k=1}^{n} x_k\right\|^2 = \sum_{k=1}^{n} \|x_k\|^2$.

5.2.6 Theorem *If $\{x_1, \ldots, x_n\}$ is an orthonormal set in the inner product space V and $x \in V$,*

$$\left\| x - \sum_{k=1}^{n} \alpha_k x_k \right\| \text{ is minimized when } \alpha_k = \langle x, x_k\rangle, \quad k = 1, \ldots, n.$$

Proof.

$$\begin{aligned}\left\| x - \sum_{k=1}^{n} \alpha_k x_k \right\|^2 &= \left\langle x - \sum_{j=1}^{n} \alpha_j x_j,\ x - \sum_{k=1}^{n} \alpha_k x_k \right\rangle \\ &= \|x\|^2 - \sum_{k=1}^{n} \alpha_k \langle x_k, x\rangle - \sum_{j=1}^{n} \bar{\alpha}_j \langle x, x_j\rangle \\ &\quad + \left\langle \sum_{j=1}^{n} \alpha_j x_j,\ \sum_{k=1}^{n} \alpha_k x_k \right\rangle.\end{aligned}$$

The last term on the right is $\sum_{j=1}^{n} |\alpha_j|^2$ since the x_k are orthonormal. Further, $-\alpha_j \langle x, x_j\rangle - \bar{\alpha}_j \langle x, x_j\rangle + |\alpha_j|^2 = |\alpha_j - \langle x, x_j\rangle|^2 - |\langle x, x_j\rangle|^2$. Thus

$$\left\| x - \sum_{k=1}^{n} \alpha_k x_k \right\|^2 = \|x\|^2 - \sum_{j=1}^{n} |\langle x, x_j\rangle|^2 + \sum_{j=1}^{n} |\alpha_j - \langle x, x_j\rangle|^2 \qquad (1)$$

so that we can do no better than to take $\alpha_j = \langle x, x_j\rangle$. ∎

The above computation establishes the following important inequality.

3.1.7 Bessel's inequality. If $\{e_n\}$ is an orthonormal subset of an inner product space X and x is an element of X, then

$$\|x\|^2 \geq \sum_{n=1}^{\infty} |\langle x, e_n \rangle|^2.$$

In other words, $\langle x, e_n \rangle = 0$ for all but countably many $n \in \mathbb{N}$, say $n = n_1, n_2, \ldots$, and

$$\|x\|^2 \geq \sum_{k=1}^{\infty} |\langle x, e_{n_k} \rangle|^2.$$

Equality holds if $\sum_{k=1}^{\infty} \langle x, e_{n_k} \rangle e_{n_k} = x$ and $n \neq n_k$.

Proof. For $N = 1, 2, \ldots$, we use $\{e_1, \ldots, e_N\}$ in Eq. (3.1.6) of 3.1.6 to obtain $0 \leq \|x - \sum_{n=1}^{N} \langle x, e_n \rangle e_n\|^2 = \|x\|^2 - \sum_{n=1}^{N} |\langle x, e_n \rangle|^2 \geq 0$. ∎

We now consider another basic geometric notion, that of projection. If M is a subspace of H and x is any vector in H, x can be resolved into a component in M and a component perpendicular to M. In other words, $x = y + z$ where $y \in M$ and z is orthogonal to every vector in M. Before generalizing to an arbitrary space we motivate some terminology.

3.1.8 Definitions. A subspace or linear manifold of a vector space L is a subset M of L that is also a vector space under the operations of L; M is closed under addition and scalar multiplication. The subset $M = \{0\}$ and the whole space L are trivial subspaces of L. If M is a subspace and is also a closed subset of L, M is a closed subspace of L.

A subset M of the vector space L is said to be convex iff $tx + (1-t)y \in M$ whenever $x, y \in M$ and $0 \leq t \leq 1$. In other words, M is convex if M contains the line segment $\{tx + (1-t)y : 0 \leq t \leq 1\}$ whenever it contains the points x and y.

A key fact is as follows: If M is a closed convex subset of the Hilbert space H and x is an arbitrary point of H, there is a unique point in M closest to x.

3.1.9 Theorem. Let M be a nonempty closed convex subset of the Hilbert space H. If $x \in H$, there is a unique element $y_0 \in M$ such that

$$\|x - y_0\| = \inf\{\|x - y\| : y \in M\}.$$

Proof. Let $d = \inf\{\|x - y\| : y \in M\}$, and pick points $y_1, y_2, \ldots \in M$ with $\|x - y_n\| \to d$ as $n \to \infty$; we show that $\{y_n\}$ is a Cauchy sequence.

Since $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$ for all $u, v \in H$ by the parallelogram law (3.1.5), we may set $u = y_n - x$, $v = x - y_m$ and obtain

$$\|y_n - y_m\|^2 + \|y_n + y_m - 2x\|^2 = 2\|y_n - x\|^2 + 2\|y_m - x\|^2$$

or

$$\|x_n - y_n\|^2 = 4\|x_n - z^{-}\|^2 \le \|x_n - z^{-}\|^2 - 4\left\|\tfrac{1}{2}(x_n - y_n) - z\right\|^2$$

Since $\frac{1}{2}(x_n - y_n) \in M$ by convexity, $\left\|\frac{1}{2}(x_n + y_n) - z\right\|^2 \ge d^2$, and it follows that $\|x_n - y_n\| \to 0$ as $n, m \to \infty$.

By completeness of M, x_n approaches a limit y_0. Hence $x_n - y_n \to y_0 - y_0 = 0$. But then $\|x_n - z\| \to d$, and $y_0 \in M$ since M is closed, this brings us to the conclusion of the proof.

To prove uniqueness, if $y_0, y_0' \in M$, with $\|x - y_0\| = \|x - y_0'\| = d$, in the parallelogram law take $u = x - y_0$, $v = x - y_0'$, to obtain

$$\|y_0 + y_0' - 2x\|^2 + \|y_0 - y_0'\|^2 = 2\|x - y_0\|^2 + 2\|x - y_0'\|^2 = 4d^2.$$

But $\left\|\frac{1}{2}(y_0 + y_0') - x\right\|^2 \ge d^2$, hence $\|y_0 - y_0'\| = 0$, so $y_0 = y_0'$. $\square$

If M is a closed subspace of H, the element y_0 found in 3.2.9 is called the projection of x on M. The following result helps us [illegible] this terminology.

3.2.10 **Theorem** Let M be a closed subspace of the Hilbert space H, and y_0 an element of M. Then

$$\|x - y_0\| = \min\{\|x - y\| : y \in M\} \quad \text{if} \quad x - y_0 \perp M,$$

that is, $x - y_0 \perp y$ for all $y \in M$.

Proof. Assume $x - y_0 \perp M$. If $y \in M$, then

$$\begin{aligned}\|x - y\|^2 &= \|x - y_0 - (y - y_0)\|^2 \\ &= \|x - y_0\|^2 + \|y - y_0\|^2 - 2\operatorname{Re}\langle x - y_0, y - y_0\rangle \\ &= \|x - y_0\|^2 + \|y - y_0\|^2 \quad \text{since } x - y_0 \perp M \\ &\ge \|x - y_0\|^2.\end{aligned}$$

Therefore, $\|x - y_0\| = \min\{\|x - y\| : y \in M\}$.

Conversely, assume $\|x - y_0\| = \min\{\|x - y\| : y \in M\}$. Let $y \in M$ and let λ be an arbitrary complex number. Then $y_0 - \lambda y \in M$ since M is a subspace, hence $\|x - y_0\| \le \|x - y_0 + \lambda y\|$. But

$$\|x - y_0 + \lambda y\|^2 = \|x - y_0\|^2 + |\lambda|^2\|y\|^2 - 2\operatorname{Re}\langle x - y_0, \lambda y\rangle.$$

Hence

$$\|x - y\|^2 = \|x\|^2 - \|y\|^2 = \|x\|^2 - \|x_1\|^2 - \|x_2\|^2 = 0.$$

Thus $x = y$, so $x \in \mathcal{M}_1 + \mathcal{M}_2$. Therefore $x = x_1 + x_2$ and $\|x\|^2 = \|x_1\|^2 + \|x_2\|^2$. Thus $x = x_1 + x_2$, $x_1 \in \mathcal{M}_1$, $x_2 \in \mathcal{M}_2$, and $\|x\|^2 = \|x_1\|^2 + \|x_2\|^2$. The representation is unique because if $x = x_1 + x_2$ is any other such decomposition, then $x_1 - x_1' = x_2' - x_2$, so $x_1 - x_1' \in \mathcal{M}_1 \cap \mathcal{M}_2 = \{0\}$. □

We may give another way of characterizing the projection of x on $\mathcal{M}$.

3.2.11 Projection Theorem. *If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, there is a unique representation $x = y + z$ with $y \in \mathcal{M}$ and $z \perp \mathcal{M}$. Furthermore, y is the projection of x on $\mathcal{M}$.*

Proof. Let y be the projection of x on $\mathcal{M}$, and let $z = x - y$. Then $x = y + z$, and by 3.2.10, $z \perp \mathcal{M}$. This proves the existence of the desired representation. To prove uniqueness, let $x = y + z = y' + z'$ where $y, y' \in \mathcal{M}$ and $z, z' \perp \mathcal{M}$. Then $y - y' \in \mathcal{M}$ since $\mathcal{M}$ is a subspace, and $y - y' \perp \mathcal{M}$ since $y - y' = z' - z$. Thus $y - y'$ is orthogonal to itself, hence $y - y' = 0$. Therefore $y = y'$, proving uniqueness.

If M is any subset of $\mathcal{H}$, the set $M^\perp = \{x \in \mathcal{H} : x \perp M\}$ is a closed subspace by definition of the inner product and 3.2.8. If M is a closed subspace, M is called the orthogonal complement of $M^\perp$, and the projection theorem is expressed by saying that $\mathcal{H}$ is the orthogonal direct sum of M and $M^\perp$, written $\mathcal{H} = M \dotplus M^\perp$.

In $\mathbb{R}^3$ it is possible to construct an orthonormal basis, that is, a set $\{y_1, \ldots, y_n\}$ of n mutually perpendicular unit vectors. Any vector x in $\mathbb{R}^3$ may then be expressed as $x = \sum_1^n (x, y_i) y_i$, so that x is a linear combination of $y_1, \ldots, y_n$. We can now also be guaranteed this fact in an arbitrary Hilbert space. The following terminology will be used.

3.2.12 Definitions. If S is a subset of the inner product space E, the *space spanned by S*, denoted by $[S]$, is the smallest closed subspace of E containing all elements of S. If $[S] = E$, S is said to *span* E. If S is finite and spans E, then $[S]$ consists of all elements $\sum_{i=1}^n \alpha_i x_i$ where $x_i \in S$, $\alpha_i \in \mathbb{C}$ $(i = 1, \ldots, n)$, $n = 1, 2, \ldots$. Then $[S] = L(S)$.

If S is a subset of the Hilbert space E, S is said to be an *orthonormal basis* for E iff S is a maximal orthonormal subset of E; in other words, S is not a proper subset of any other orthonormal subset of E. An orthonormal set $S \subseteq E$ is maximal iff $[S] = E$, and there are several other conditions equivalent to this, as we now prove.

3.2.13 Theorem. Let $B = \{x_\alpha : \alpha \in I\}$ be an orthonormal subset of the Hilbert space H. The following conditions are equivalent:

(a) B is an orthonormal basis.

(b) B is a "complete orthonormal set," that is, the only $x \in H$ such that $x \perp B$ is $x = 0$.

(c) B spans H, that is, $\overline{\operatorname{span}}(B) = H$.

(d) For every $x \in H$, $x = \sum_{\alpha \in I} x_\alpha \langle x, x_\alpha \rangle$ (only a countable number of terms are nonzero). By 3.2.12, $\langle x, x_\alpha \rangle = 0$ for all but countably many x_α, say for $\alpha = \alpha_1, \alpha_2, \ldots$; the assertion is that $\sum_{n=1}^{\infty} \langle x, x_{\alpha_n} \rangle x_{\alpha_n} = x$, and this holds regardless of the order in which the x_{α_n} are listed.

(e) For all $x, y \in H$, $\langle x, y \rangle = \sum_{\alpha \in I} \langle x, x_\alpha \rangle \langle x_\alpha, y \rangle$.

(f) For all $x \in H$, $\|x\|^2 = \sum_{\alpha \in I} |\langle x, x_\alpha \rangle|^2$.

Condition (f) [and sometimes (b) as well] is referred to as the Parseval relation.

Proof. (a) $\Rightarrow$ (b). Suppose that $x \perp B$, $x \neq 0$, say $\|x\| = r > 0$. Then $B \cup \{x/r\}$ is an orthonormal set, contradicting the maximality of B.

(b) $\Rightarrow$ (c). If $x \in H$, write $x = y + z$ where $y \in S(B)^-$ and $z \perp S(B)^-$ (see 3.2.11). By (b), $z = 0$ since $z \perp S(B)$.

(c) $\Rightarrow$ (d). Since $S(B) = \overline{S(B)}$, given $x \in H$ and $\varepsilon > 0$ there is a finite set $F \subset I$ and complex numbers α_β, $\beta \in F$, such that

$$\left\| x - \sum_{\beta \in F} \alpha_\beta x_\beta \right\| < \varepsilon.$$

By 3.2.6, if G is any finite subset of I such that $F \subset G$,

$$\left\| \sum_{\beta \in G} \langle x, x_\beta \rangle x_\beta - x \right\| \leq \left\| \sum_{\beta \in G} \alpha_\beta x_\beta - x \right\| \quad \text{where } \alpha_\beta = 0 \text{ for } \beta \notin F$$

$$< \varepsilon.$$

We may assume that $\langle x, x_\alpha \rangle \neq 0$ for every $\alpha \in I$.

Thus if $x_{\alpha_1}, x_{\alpha_2}, \ldots$ is any ordering of the terms $x_\alpha \in B$ for which $\langle x, x_\alpha \rangle \neq 0$, $\|x - \sum_{n=1}^{N} \langle x, x_{\alpha_n} \rangle x_{\alpha_n}\| < \varepsilon$ for sufficiently large N, as desired.

(d) $\Rightarrow$ (e). This is immediate from 3.2.12.

(e) $\Rightarrow$ (f). Set $x = y$ in (e).

(f) $\Rightarrow$ (b). If $x \perp B$, let C be an orthonormal set with $B \subset C$, $B \neq C$. If $x \neq 1 \cdot x_\beta \in B$, we have $\|x\|^2 = \sum_{\alpha \in C} |\langle x, x_\alpha \rangle|^2 = 0$, since by maximality of

$G \cdot x$ is orthogonal to everything in E. This is a contradiction because $\|x - x\| = 1$ for all $x \in G$. $\square$

3.2.34 *Corollary.* Let $S = \{x_\alpha : \alpha \in A\}$ be an orthonormal subset of H, and let $E = \overline{\operatorname{span}}(S)$.

(a) S is an orthonormal basis for E. (Note that E is a closed subspace of H and hence is itself a Hilbert space with the same inner product.)

(b) If $x \in H$ and $y = \operatorname{proj}_E x$ is the projection of x on E, then

$$y = \sum_{\alpha} \langle x, x_\alpha \rangle x_\alpha$$

[see 3.2.33(d) for the interpretation of the series].

Proof. (a) Note that the space spanned by S in E is $E \cap S^\perp = \{0\}$.

(b) By part (a) and 3.2.33, $\sum_\alpha |\langle x, x_\alpha \rangle|^2 \le \|x\|^2 < \infty$. But $x - y \perp E$ by 3.2.11, hence $\langle x, x_\alpha \rangle = \langle y, x_\alpha \rangle$ for all α. $\square$

An additional application of Zorn's lemma shows that every Hilbert space has an orthonormal basis; an additional argument shows that any two orthonormal bases have the same cardinality (see Problem 5). This fact may be used to classify all possible Hilbert spaces, as follows.

3.2.35 *Theorem.* Let S be an arbitrary set, and let H be a Hilbert space with an orthonormal basis B having the same cardinality as S. Then there is an isometric isomorphism of H onto the Hilbert space $l^2(S)$.

Proof. We may write $B = \{x_\alpha : \alpha \in A\}$. If $x \in H$, $\sum_\alpha |\langle x, x_\alpha \rangle|^2 < \infty$ by 3.2.33(b), and hence $x = \sum_{\alpha \in A} \langle x, x_\alpha \rangle x_\alpha$, where $\sum_{\alpha \in A} |\langle x, x_\alpha \rangle x_\alpha|^2 = \|x\|^2$ by 3.2.33(c). The map $x \to \{\langle x, x_\alpha \rangle : \alpha \in A\}$ is one to one from H into $l^2(S)$ by uniqueness of coefficients, and it is also linear. It must be onto. To show that the map is onto, consider any collection of complex numbers $\{c_\alpha : \alpha \in A\}$ with $\sum_\alpha |c_\alpha|^2 < \infty$. Say $c_\alpha = 0$ except for $\alpha = \alpha_1, \alpha_2, \ldots$, and let $y = \sum_n c_{\alpha_n} x_{\alpha_n}$. The series converges to an element of H because of the Riesz-Fischer fact which makes it converge fast enough to be actual sequences. If $\{y_1, y_2, \ldots\}$ is an enumeration of the values of A, the series $\sum_j c_{\alpha_j} x_{\alpha_j}$ converges to some element of H, let $\sum_j c_{\alpha_j} x_{\alpha_j} = y$, say. To see this, observe that $\|\sum_{j=m}^{n} c_{\alpha_j} x_{\alpha_j}\|^2 = \sum_{j=m}^{n} c_{\alpha_j}^2 \|x_{\alpha_j}\|^2 = \sum_{j=m}^{n} c_{\alpha_j}^2$, thus the partial sums form a Cauchy sequence in H since $\sum |c_{\alpha_j}|^2 < \infty$.

Since $\langle x_{\alpha_j}, x_{\alpha_k} \rangle = 0$ if $j \ne k$, it follows that $\langle y, x_{\alpha_j} \rangle = c_{\alpha_j}$ for all j, so that y maps onto $\{c_\alpha : \alpha \in A\}$. $\square$

Theorem 5.2.5 is not as useful as it looks. For example, when working in $\ell^2(\mathbb{N})$, we usually like advantages of what we know about $\mathbb{R}^n$.

We may also discover Hilbert spaces that are separable, that is, have a countable dense set.

5.2.26 *Theorem.* *A Hilbert space H is separable if and only if it has a countable or denumerable basis. If H is n-dimensional, then it has n elements, H is isomorphic to $\mathbb{C}^n$; if H is infinite-dimensional, then it has a countably infinite basis, H is isomorphic to $\ell^2(\mathbb{N})$, and H is unitarily equivalent to $\ell^2(\mathbb{N})$.*

Proof. Let H be an orthonormal basis for H. Now $\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\,\mathrm{Re}\langle x, y\rangle \neq \|x\|^2 + \|y\|^2$ if $x \neq y$, because the basis $\{e_\alpha = \frac{1}{\sqrt{2}}(x+y) : \alpha \in B\}$ is infinite. If D is dense in H, D must contain a point in each $\{e_\alpha : \alpha \in B\}$, so that B is countable. If D is countable, then H cannot be uncountable, for then H cannot be separable.

Now assume H is a countable set $\{x_1, x_2, \ldots\}$. If U is a nonempty open subset of $\{\|x\| = \alpha\} \cap \{\|x-y\| = \beta\}$, it contains an element of the form $\sum_{k=1}^{n} a_k x_k$ with the $a_k \in \mathbb{Q}$ an immediate consequence, may be assumed to be rational, in other words to have rational numbers as real and imaginary parts. Thus

$$D = \left\{ \sum_{k=1}^{n} a_k x_k : n = 1, 2, \ldots, \text{ the } a_k \text{ rational} \right\}$$

is a countable dense set, so that H is separable. The remaining statements of the theorem follow from 5.2.17. ∎

A linear isomorphism from one Hilbert space to another which preserves inner products is called a unitary transformation. This is a consequence of the following proposition.

5.2.27 *Polarization Identity.* *In any inner product space,*

$$4\langle x, y\rangle = \|x+y\|^2 - \|x-y\|^2 - i\|x+iy\|^2 + i\|x-iy\|^2$$

Proof.

$$\begin{aligned}
\|x+y\|^2 &= \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\langle x, y\rangle \\
\|x-y\|^2 &= \|x\|^2 + \|y\|^2 - 2\,\mathrm{Re}\langle x, y\rangle \\
\|x+iy\|^2 &= \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\langle x, iy\rangle \\
\|x-iy\|^2 &= \|x\|^2 + \|y\|^2 - 2\,\mathrm{Re}\langle x, iy\rangle
\end{aligned}$$

But $\mathrm{Re}\langle x, iy\rangle = \mathrm{Re}[-i\langle x, y\rangle] = \mathrm{Im}\langle x, y\rangle$, and the result follows. ∎

Problems

1. Let $\mathcal{H}$ be a Hilbert space. Show that the elements e_n, $n = 0, \pm 1, \pm 2, \ldots$, form an orthonormal basis, where

$$e_n(x) = \begin{cases} 0, & x \neq n, \\ 1, & x = n. \end{cases}$$

2. (a) If M is an arbitrary subset of the Hilbert space $\mathcal{H}$, show that $M^{\perp\perp} = S(M)$.

   (b) If M is a linear manifold of $\mathcal{H}$, show that M is dense in $\mathcal{H}$ if and only if $M^{\perp} = \{0\}$.

3. Let $x_1, \ldots, x_n$ be elements of a Hilbert space. Show that the x_i are linearly dependent if and only if the Gramian of the determinant of the inner products (x_i, x_j), $i, j = 1, \ldots, n$, is 0.

4. (Gram–Schmidt process) Let $A = \{x_1, x_2, \ldots\}$ be a countable linearly independent subset of the Hilbert space $\mathcal{H}$. Define $e_1 = x_1/\|x_1\|$, having chosen orthonormal elements $e_1, \ldots, e_n$ for $x_1, \ldots, x_n$, let the projection of x_{n+1} on the vector space spanned by $e_1, \ldots, e_n$ be

$$z_n = \sum_{i=1}^{n} (x_{n+1}, e_i) e_i.$$

   Define

$$e_{n+1} = \frac{x_{n+1} - z_n}{\|x_{n+1} - z_n\|}.$$

   (a) Show that $L\{e_1, \ldots, e_n\} = L\{x_1, \ldots, x_n\}$ for $n = 1, 2, \ldots$, where $z_n \neq x_{n+1}$, and the process is well defined.

   (b) Show that the e_n form an orthonormal basis for $S(A)$.

   Comments. Consider the space $\mathcal{H} = L^2[-1, 1]$ and the basis $x_n(t) = t^n$, $n = 0, 1, \ldots$, of Haar–Schmidt. Because $x_n(t) = x_n(-t)e^{in\pi}$ the Gram–Schmidt process yields $e_n(t) = e_n(-t)e^{in\pi}$, where e_n is chosen so that $\|e_n\| = 1$. Since e_n is in $L^2[-n, n]$, only we have $e_n = e^{in\pi/2}\varphi_n$, $n = 0, 1, \ldots$, where φ_n is in the *Hermite polynomials* $\varphi_n(t) = e^{t^2/2}(d^n/dt^n)e^{-t^2}$ [see].

5. (a) If you are familiar with Zorn's Lemma, show that every Hilbert space has an orthonormal basis.

   (b) If you are familiar with cardinal arithmetic, show that any two orthonormal bases have the same cardinality.

6. Let D be an open subset of the complex plane, and let $H^2(D)$ be the collection of all functions f analytic on D such that

$$\|f\| = \int_0^{2\pi} \int_0^{2\pi} |f(e^{i\theta}, e^{i\varphi})| \, d\theta \, d\varphi < \infty.$$

If we define

$$(f * g)(e^{i\theta}) = \int_0^{2\pi} f(e^{i\theta}, e^{i(\theta-\varphi)}) g(e^{i(\theta-\varphi)}) \, d\varphi \qquad (g \in A(\mathbb{D})),$$

$A(\mathbb{D})$ becomes an integral domain.

(b) If K is a compact subset of $\mathbb{D}$ and $f \in A(\Omega)$, show that

$$\sup\{|f(z)| : z \in K\} \leq \|f\| / 2\pi\delta,$$

where δ is the distance from K to the complement of Ω. Therefore convergence in $A(\Omega)$ implies uniform convergence on compact subsets of Ω. If $|a| < R$, the Cauchy integral formula yields

$$f(a) = (2\pi i)^{-1} \int_{|z|=R} f(z)(z-a)^{-1} dz, \qquad r < R < \rho.$$

Integrate this equation with respect to θ and φ as in (a). Now observe that if Ω is the entire plane, we may take $\rho = \infty$, and it follows that $A(\mathbb{C}) = \{0\}$.

(c) Show that $A(\mathbb{D})$ is complete and therefore is a Hilbert space.

7. (a) If f is analytic on the unit disk $\mathbb{D} = \{z : |z| < 1\}$ with Taylor expansion $f(z) = \sum_{n=0}^{\infty} a_n z^n$, show that

$$\sup_{0<r<1} \frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta = \sum_{n=0}^{\infty} |a_n|^2.$$

In addition, $\sup_{0<r<1} \int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta < \infty$ if and only if f is in the Hardy space

$$H^2(\mathbb{D}) = \sup_{0<r<1} \frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta < \infty.$$

Then H^2, with norm $\|\cdot\|_2$, is a pre-Hilbert space.

(b) If $f \in H^2$, show that

$$\int_0^{2\pi} |f(re^{i\theta}) - f(e^{i\theta})|^2 \, d\theta \to 0 \text{ as } r \to 1^-.$$

Hence $H^2 \subset A(\mathbb{D})$, and convergence in H^2 implies convergence in $A(\mathbb{D})$.

(c) If $f_n(x) \to f(x)$, $n = 0, \pm 1, \pm 2, \ldots$, show that $f_n \to f$ in $W^2[0,1]$ but not in $L^2[0,1]$.

(d) Show that W^k is complete, and hence is a Hilbert space. (By fact, W^k is isometrically isomorphic to a subspace of L^2.

(e) If $f_n(x) = e^{inx}$, $n = 0, \pm 1, \pm 2, \ldots$, show that the f_n form an orthonormal basis for W^k.

(f) If $f_n(x) = [(n^2+1)/2\pi]^{1/2} e^{-|n|x}$, $n = 0, \pm 1, \pm 2, \ldots$, show that the f_n form an orthonormal basis for $W^1[0,\infty)$.

8. Let $\mathcal{M}$ be a closed convex subset of the Hilbert space H, and y an element of H. If $x \in \mathcal{M}$, show that

$$\|y - x\| = \inf\{\|y - z\| : z \in \mathcal{M}\}$$

iff

$$\mathrm{Re}\langle y - x, z - x\rangle \le 0 \quad \text{for all } z \in \mathcal{M}.$$

9. (a) If α is a continuous complex-valued function on $[0, 2\pi]$ with $\alpha(0) = \alpha(2\pi)$, use the Stone–Weierstrass theorem to show that α can be uniformly approximated by a trigonometric polynomial $\sum_{n=-N}^{N} a_n e^{inx}$. Conclude that any 2π-periodic polynomial can be found in $L^2[0, 2\pi]$.

(b) If $f \in L^2[0, 2\pi]$, show that the Fourier series $\sum_{n=-\infty}^{\infty} a_n e^{inx}$, $a_n = (1/2\pi)\int_0^{2\pi} f(x)e^{-inx}\,dx$, converges to f in L^2, that is,

$$\int_0^{2\pi} |f(x) - \sum_{n=-N}^{N} a_n e^{inx}|^2\,dx \to 0 \quad \text{as} \quad N \to \infty.$$

(c) Show that $e^{inx}/\sqrt{2\pi}$, $n = 0, \pm 1, \pm 2, \ldots$, yields an orthonormal basis for $L^2[0, 2\pi]$.

10. (a) Give an example to show that if $\mathcal{M}$ is a nonempty, closed, but not convex subset of a Hilbert space H, there need not be an element of minimum norm in $\mathcal{M}$. (Thus the convexity hypothesis cannot be dropped from Theorem 3.2.3, even if we restrict ourselves to subspaces and forget about uniqueness.)

(b) Show that convexity is not necessary in the conclusion of Exercise 3.2.9 if $\mathcal{M}$ is finite-dimensional.

3.3 Linear Transformations; Adjoints; Linear Spaces

The idea of a linear transformation from one Euclidean space to another is familiar. If A is a 3×2 matrix from $\mathbb{R}^2$ to $\mathbb{R}^3$, then A is completely specified

by placing us rather on a basis of ... for ... $\sum_{i=1}^{n} \ldots$ $\sum_{i=1}^{n}$ ... finite-dimensional spaces always coincide. It expresses ... and ... the uniqueness of the ... vectors. It is represented by ... matrix. If ... so that A is a linear transformation on $\mathbb{R}^n$, A is one-to-one iff it is onto and A^{-1} ... is always continuous (as well as linear).

Linear transformations on infinite-dimensional spaces have many features not found in the finite-dimensional case, as we shall see.

In this section, we study mappings A from one normed linear space X to another such space Y. The mapping A will be a linear operator, that is, $A(\alpha x + \beta y) = \alpha(Ax) + \beta(Ay)$ for all $x, y \in X$ and $\alpha, \beta \in \mathbb{K}$. We use the symbol $\|\cdot\|$ for the norm on both spaces, as confusion should not result. Linear operators can of course be defined on arbitrary vector spaces, but in this section it is always understood that the domain and range are normed.

Everything does not apply everywhere to study, as ... we encounter a new concept.

3.3.1 Definition and Comments. If A is a linear operator, the norm of A is defined by

$$\|A\| = \sup\{\|Ax\| : x \in X, \|x\| \le 1\}.$$

We may express $\|A\|$ in two other ways:

$$\|A\| = \sup\{\|Ax\| : x \in X, \|x\| = 1\},$$

$$\|A\| = \sup\{\|Ax\|/\|x\| : x \in X, x \neq 0\}.$$

... note this, note that $\|Ax\| = \|A(x/\|x\|)\|$ ... if $x \neq 0$, then $\|Ax\| = \|A\|\,\|x\|$. ... and so ... let ... hence ... $\le$ ... Finally if $\|x\| \le 1$ ... then $\|Ax\| \le \|Ax\|$, so $\|A\| \le \ldots$)

It follows from (2) that $\|Ax\| \le \|A\|\,\|x\|$, and in fact $\|A\|$ is the smallest number k such that $\|Ax\| \le k\|x\|$ for all $x \in X$.

The linear operator A is said to be *bounded* if $\|A\| < \infty$. Boundedness is often easy to check, and it is fortunate circumstance because we can show that boundedness is equivalent to continuity.

3.3.2 Theorem. *A linear operator A is continuous iff it is bounded.*

Proof. If A is bounded and $\{x_n\}$ is a sequence in X converging to 0, then $\|Ax_n\| \le \|A\|\,\|x_n\| \to 0$. Thus $\|Ax_n\| < \varepsilon$ for all n ... the mapping $x \to \|x\|$ of X into the nonnegative reals is continuous; thus follows because $|\|x\| - \|y\|| \le \|x - y\|$. Thus A is continuous at 0, and therefore, by linearity, it is continuous everywhere. On the other hand, if A is unbounded, we can find elements $x_n \in X$ with $\|x_n\| = 1$ and $\|Ax_n\| \to \infty$. Let $y_n = x_n / \|Ax_n\|$; then $y_n \to 0$, but $\|Ay_n\| = 1$ for all n, hence Ay_n does not converge to 0. Consequently, A is discontinuous. ...

We are going to show that the set of all bounded linear operators from X to X is a normed linear space, but first we consider some examples.

1.5.1 Example (a) Let $X = C[a, b]$ the set of all continuous complex-valued functions on the closed bounded interval $[a, b]$ or reals. For the norm on X

$$\|x\| = \sup\{|x(t)| : a \le t \le b\}, \qquad x \in C[a, b].$$

With this norm, X is a Banach space (see 1.2.2(a)). Let $k = k(s, t)$ be a continuous complex-valued function on $[a, b] \times [a, b]$, and define a linear operator on X by

$$(Ax)(s) = \int_a^b k(s, t)x(t)\,dt, \qquad a \le s \le b.$$

(By the dominated convergence theorem, Ax actually belongs to X.) We show that A is bounded:

$$\begin{aligned}\|Ax\| &= \max_{a \le s \le b} \left| \int_a^b k(s, t)x(t)\,dt \right| \\ &\le \max_{a \le s \le b} \|x\| \max_{a \le s \le b} \int_a^b |k(s, t)|\,dt \\ &= \|x\| \max_{a \le s \le b} \int_a^b |k(s, t)|\,dt\end{aligned}$$

since $\int_a^b |k(s, t)|\,dt$ is a continuous function of s by the dominated convergence theorem. Thus

$$\|A\| \le \max_{a \le s \le b} \int_a^b |k(s, t)|\,dt < \infty.$$

In fact $\|A\| = \max_{a \le s \le b} \int_a^b |k(s, t)|\,dt$ (see Problem 5).

(b) Let $X = l^p(S)$, where S is the set of all integers and $1 \le p < \infty$. Define a linear operator T on X by $(Tx)_n = x_{n-1}$, $n \in S$. T is called the *unilateral shift* or the *backward shift*. It follows from the definition that T is one-to-one, onto, and $(T^{-1}y)_n = y_{n+1}$, $n \in S$. Again $\|Tx\| = \|x\|$ for all $x \in X$; hence $\|T^{-1}y\| = \|y\|$ for all $y \in X$, so that T is an isometric automorphism of X with itself. In particular, $\|T\| = \|T^{-1}\| = 1$.

If we replace S by the positive integers and define T as above, the resulting operator is called the *unilateral* or *unilateral shift*. Let us call the shift in (b) the *bilateral shift*. It is the unique operator T such that $T(x_1, x_2, \ldots) = (y_1, y_2, \ldots)$,

like shift operators we have defined the shifts to the left. We may also define shifts to the right. In the two sided case we have $Sf(j) = f(j-1)$, $j \in \mathbb{Z}$ and $S = T^{-1}$. In the one-sided case we set $Sf(0) = 0$, $Sf(j) = f(j-1)$, $j \geq 1$, $f \in l^2(\mathbb{N})$. Then $S(f_0, f_1, f_2, \ldots) = (0, f_0, f_1, \ldots)$. The operator S is one-to-one and its range $S(l^2)$ is the closed subspace of l^2 consisting of those sequences whose first coordinate is 0.

(c) Assume that the Banach space X is the direct sum of the two closed subspaces M and N, in other words each $x \in X$ can be represented in a unique way as $x = y + z$ for some $y \in M$, $z \in N$. (We have already encountered this situation with the Hilbert space H and closed subspace S — $H = S + S^{\perp}$.) We define a linear operator P on X by

$$Px = y.$$

P is called a projection; specifically P is the projection of X on M and it has the following properties:

(1) P is idempotent that is $P^2 = P$, where P^2 is the composition of P with itself.

(2) P is continuous.

Property (1) follows from the reduction of P; property (2) will be proved later as a consequence of the closed graph theorem (see also 3.4.1 b).

Conversely, let P be a continuous idempotent linear operator on X. Define

$$M = \{x \in X : Px = x\}, \qquad N = \{x = Px : x \in X\}.$$

Then we can show that M and N are closed subspaces, X is the direct sum of M and N, and P is the projection of X on M.

By continuity of P, M and N are closed subspaces. If $x \in N$, then $x = Px = P(Px) = Py$ for some $y \in X$. Since $P(I-P) = (I-P)P = 0$, $I-P$ is the direct sum of M and N by construction. Thus $x = y$ by definition of y, so that P is the projection of X on M.

If y is a linear mapping from a vector space X to the scalar field, y is called a linear functional. The study of a space X is helped by the [illegible] insight is gained about X and linear spaces by studying ways of representing continuous linear functionals on such spaces. We give some examples.

3.2.4 *Representations of Continuous Linear Functionals.* (a) Let y be a continuous linear functional on the Hilbert space H. We show that there is a unique element $y \in H$ such that

$$f(x) = (x, y) \qquad \text{for } x \in H, \quad y \in H.$$

This is one of several results we will call the Riesz representation theorem.

In the desired inequality must be unique, or $\|x\| = \|y\|$ for all x then $x = y$. Conjecture to everything in N (including itself), so $y = z$.

To prove existence, let N be the null space of f, that is, $N = \{x \in H : f(x) = 0\}$. If $N = H$, then $f = 0$, and we may take $y = 0$. Otherwise, we have an element $w \in N^\perp$ with $w \neq 0$. Then $w \notin N$, and if we define $v = w/\|w\|$, we have $v \in N^\perp$ and $\|v\| = 1$.

If $x \in H$ and $f(x) = \alpha$, then

$$x = (x - \alpha v) + \alpha v \qquad \text{with} \qquad x - \alpha v \in N, \quad \alpha v \in N^\perp.$$

If $y = \bar{\alpha} v / \|v\|^2$, then

$$\begin{aligned} \langle x, y \rangle &= \langle x - \alpha v, y \rangle + \langle \alpha v, y \rangle \\ &= \alpha \langle v, y \rangle \qquad \text{since} \qquad y \perp N \\ &= \alpha = f(x) \end{aligned}$$

as desired.

The above argument shows that if f is not identically 0, then $N^\perp$ is one-dimensional. For if $x \in N^\perp$ and $f(x) = \alpha$, then $x - \alpha v \in N \cap N^\perp = \{0\}$, hence $x = \alpha v$. The space $N = \{x : f(x) = 0\}$.

Notice also that $\alpha = \|f\|$ is the norm of f considered as a linear operator, for

$$|f(x)| = |\langle x, y \rangle| \leq \|x\| \, \|y\| = \|x\| \, \|f\|,$$

so $\|f\| \leq \|y\|$; but $|f(y)| = |\langle y, y \rangle| = \|y\|^2 = \|y\| \, \|f\|$, so $\|f\| = \|y\|$.

Now consider the space H^* of all continuous linear functionals on H. H^* is a vector space under the usual operations of addition and scalar multiplication. If $y \in H$, we associate the functional f_y in H^* given by the Riesz representation theorem; we obtain a map $\varphi : H \to H^*$ that is one-to-one, onto, norm-preserving, and conjugate linear, that is,

$$\varphi(\alpha f + \beta g) = \bar{\alpha} \varphi(f) + \bar{\beta} \varphi(g), \qquad f, g \in H^*, \quad \alpha, \beta \in \mathbb{C}.$$

[Note that if $f(x) = \langle x, y \rangle$ for all x, then $\varphi(f) = \{x, y\}$.] Such a map is called a conjugate isometry.

(b) Let f be a continuous linear functional on $H^* = H^*(\mathbb{R})$, where H is the set of positive integers. For $x \in l^2$, we know that if x is defined by $(Tx)(n) = \frac{1}{n} x(n)$, then T is a unique element $y = (y_1, y_2, \ldots) \in l^2$ such that

$$f(x) = \sum_{n=1}^{\infty} x_n y_n \qquad \text{for all} \qquad x \in l^2$$

Furthermore,

$$\|f\| = \|y\| = \left(\sum_{k=1}^{\infty} |\eta_k|^2\right)^{1/2}.$$

To prove this, we let $e^{(j)}$ be the sequence in l^2 defined by $e^{(j)}(j) = 0$, $j \neq k$; $e^{(j)}(j) = 1$. If $x \in l^2$, then $x = \sum_{j=1}^{\infty} \alpha_j e^{(j)}$ where $\alpha_j = \xi_j$. Hence $y = \sum_{j=1}^{\infty} \xi_j \eta_j e^{(j)}$, where the series converges in l^2 by assumption on y.

$$f(x) = \sum_{j=1}^{\infty} \xi_j \eta_j, \qquad \text{where} \qquad x = \{\xi_j\} \in l^2. \tag{1}$$

Now write y in polar form $\eta_j = |\eta_j| e^{i\theta_j}$, $\eta_j \neq 0$. Let

$$z = \{\zeta_j\} = \{e^{-i\theta_j}\}, \qquad \text{so} \qquad \zeta_j \eta_j = |\eta_j|, \quad j = 1, 2, \cdots.$$

By (1),

$$f(z) = \sum_{j=1}^{\infty} \zeta_j \eta_j = \sum_{j=1}^{\infty} |\eta_j|. \tag{2}$$

But

$$\begin{aligned} |f(z)| &\leq \|f\| \, \|z\| \\ &= \|f\| \left(\sum_{j=1}^{\infty} |e^{-i\theta_j}|^2\right)^{1/2} \\ &= \|f\| \left(\sum_{j=1}^{\infty} 1^2\right)^{1/2}. \end{aligned}$$

By Eq. (2),

$$\left(\sum_{j=1}^{\infty} 1^2\right)^{1/2} \leq \|f\|.$$

Hence $y \in l^2$ and $\|y\| = \|f\|$. To prove that $\|f\| \leq \|y\|$, observe that $\|f\|$ is the least M such that $|f(x)| \leq M\|x\|$ for all $x \in l^2$. Since $\|x\|$ is a positive integer and μ is counting measure, by Hölder's inequality, $|f(x)| \leq \|y\| \, \|x\|$, so $\|f\| \leq \|y\|$.

Finally, we prove uniqueness. If $y, z \in l^2$ and $f(x) = \sum_{j=1}^{\infty} \xi_j \eta_j = \sum_{j=1}^{\infty} \xi_j \zeta_j$ for all $x \in l^2$, then $g(x) = \sum_{j=1}^{\infty} \xi_j (\eta_j - \zeta_j) = 0$ for all $x \in l^2$.

The above equation with y replaced by x shows that $\|x\| = \|x - x_0\|$, hence $\|x\| = \|x - x_0\| = 0$, and thus $x = x_0$.

(b) Let f be a continuous linear functional on X. We show that there is a unique element $y \in X^*$ such that

$$f(x) = \sum_{j=1}^{\infty} x_j y_j \qquad \text{for all} \qquad x = (x_j) \in X.$$

Furthermore $\|f\| = \|y\| = \sup_j |y_j|$.

The argument of the above may be repeated up to Eq. (1). In this case, however, if $x_j = \alpha e^{(j)}$, $j = 1, 2, \ldots$, we take

$$z_j = |\beta| \qquad \text{if } \alpha^{(j)} \neq 0, \qquad z_j = \beta \qquad \text{if } \alpha^{(j)} = 0 \text{ in position } n.$$

Then by (1),

$$f(x_j) = \alpha^{(j)} z_j \alpha^{(j)} = |\alpha_j|.$$

But

$$|f(x_j)| \leq \|f\| \, \|x_j\| = \|f\| \, |\alpha_j|$$

therefore $|\alpha_j| \leq \|f\|$ for all n, so that $y \in l^\infty$ and $\|y\| \leq \|f\|$. But by (1),

$$|f(x)| \leq \left(\sup_j |y_j|\right) \sum_{j=1}^{\infty} |\alpha_j| = \|y\| \, \|x\|$$

hence $\|f\| \leq \|y\|$. Uniqueness is proved as in (b).

(c) Still with y_j fixed, the map $T : x \mapsto (y_j x_j)$ of l^p into l^p is linear and norm-preserving, hence one-to-one. To show that it is onto, observe that any linear functional on the form $f(x) = \sum_{j=1}^{\infty} y_j x_j$, $x \in l^p$, $y \in l^q$, satisfies $\|f\| \leq \|y\|$ for the analysis of 5.3 still applies. Therefore T is continuous, so that every $y \in l^q$ is the image of some $f \in (l^p)^*$. Since $\|T\| = \|y\|$, we have an isometric isomorphism of $(l^p)^*$ onto l^q.

If we replace y_j by T_j, we obtain the result that there is a unique $x \in l^p$ such that $f(x) = \sum_{j=1}^{\infty} x_j y_j$ for all $y \in l^q$. This makes the map $f \mapsto x$ a conjugate isometry rather than an isometric isomorphism. If $p = 2$ then $q = 2$ as well, and thus we have another proof of the Riesz representation theorem for separable Hilbert spaces (see 3.3.6(c)). In fact, essentially the same argument may be used in any Banach space X that is uniformly convex (as explained by an exercise in Banach spaces). Other examples of representations of continuous linear functionals are given in Problems 10 and 11.

We now show that the set of bounded linear operators from one normed linear space to another can be made into a normed linear space.

3.3.5 *Theorem.* Let L and M be normed linear spaces and let $\mathcal{B}(L, M)$ be the collection of all bounded linear operators from L to M. The operator norm defined by (3.3.1) is a norm on $\mathcal{B}(L, M)$ and, if M is complete, then $\mathcal{B}(L, M)$ is complete. In particular, the set L^* of all continuous linear functionals on L is a Banach space (whether or not L is complete).

Proof. It follows from (3.3.1) that $\|A\| \geq 0$, and $\|\alpha A\| = |\alpha| \, \|A\|$ for all $\alpha \in \mathbb{C}$ and $A \in \mathcal{B}(L, M)$. Also by (3.3.1), $\|A\| = 0$ only if $Ax = 0$ for all $x \in L$, hence $A = 0$. If $A, B \in \mathcal{B}(L, M)$, then again by (3.3.1),

$$\|A + B\| = \sup\{\|Ax + Bx\| : x \in L, \|x\| \leq 1\}$$

$$\leq \sup\{\|Ax\| + \|Bx\| : x \in L, \|x\| \leq 1\}$$

$$\leq \|A\| + \|B\|.$$

Since $\|AB\| \leq \|A\| \, \|B\|$ for all A, B, we have

$$\|A - B\| \geq \|A\| - \|B\|$$

and it follows that $\mathcal{B}(L, M)$ is a normed space. If the space M is complete, then $\mathcal{B}(L, M)$ is in fact a Banach algebra.

Now let $(A_1, A_2, \ldots)$ be a Cauchy sequence in $\mathcal{B}(L, M)$. Then

$$\|A_n x - A_m x\| \leq \|A_n - A_m\| \, \|x\| < \varepsilon \|x\| \quad \text{for } n, m > N. \tag{*}$$

Therefore $(A_n x)$ is a Cauchy sequence in M for each $x \in L$, hence $A_n x$ converges pointwise on L to an operator A. Since the A_n are linear, so is A. Observe that $A_n x - Ax = \lim_{m \to \infty} (A_n x - A_m x)$, and letting $m \to \infty$ in (*), we have $\|A_n x - Ax\| \leq \varepsilon \|x\|$ for $n > N$. Now choose $n > N$ such that $\|A_n - A\| < \varepsilon$. For $x \neq 0$ in L, we have $\|A_n x - Ax\| < \varepsilon \|x\|$, or equivalently, $\|(A_n - A)x\| < \varepsilon \|x\|$. Therefore $\|A_n - A\| \leq \varepsilon$. Thus $\|A_n - A\| \to 0$ as $n \to \infty$. Since $\|A\| \leq \|A - A_n\| + \|A_n\|$, we have $A \in \mathcal{B}(L, M)$ and $A_n \to A$ in the operator norm. $\square$

In the above proof we have used three different types of convergence of sequences of operators.

3.3.6 *Definitions and Comments.* Let $A, A_1, A_2, \ldots \in \mathcal{B}(L, M)$. We say that A_n converges *uniformly* to A if $\|A_n - A\| \to 0$ (equivalently, $A_n \to A$ in $\mathcal{B}(L, M)$). Since $\|(A_n - A)x\| \leq \|A_n - A\| \, \|x\|$, uniform operator convergence means that $A_n x \to Ax$ uniformly for $\|x\| \leq 1$ (or equivalently on $\|x\| = 1$, or $\|x\| < 1$, or any positive real number).

We say that A_n converges *pointwise* to A if $A_n x \to Ax$ for each $x \in L$. Thus pointwise operator convergence is precisely convergence in the sense of §1.

Uniform convergence implies strong convergence, but not conversely. For example, let $(e_1, e_2, \ldots)$ be an orthonormal set in a Hilbert space, and let

$x_n(t) = \sin nt$, $n = 1, 2, \ldots$. Then x_n converges pointwise to 0 by Bessel's inequality but x_n does not converge uniformly to 0. In fact $\|x_n\| = 1$ for all n, hence $\|x_n\| \not\to 0$.

There is an important principle of finite-dimensional spaces that we are now in a position to discuss. If the pervading assumption is replaced by that of a Hilbert space, so that $X = \mathbb{C}^N$, the norm of a vector x was taken to be the Euclidean norm $\|x\| = \left(\sum_{i=1}^N |x_i|^2\right)^{1/2}$. The matrix associated with this norm yields the standard operator norm on $\mathbb{C}^N$. However we may put numerous other norms on $\mathbb{C}^N$, for example, the p-norm $\|x\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{1/p}$, $1 \le p < \infty$, or the ∞-norm $\|x\|_\infty = \max\{|x_1|, \ldots, |x_N|\}$. Since the space is finite dimensional these norms cause no difficulties and all elements have finite norms. Furthermore, the introduction of norms causes no problems because all norms give rise to the same topology, which is the usual topology on $\mathbb{C}^N$. The proof of this result is outlined in Problems 6 and 7.

Problems

1. Let f be a linear functional on the normed linear space X. If f is not identically 0, show that the following are equivalent:
   (a) f is continuous.
   (b) The null space $N = f^{-1}\{0\}$ is closed.
   (c) N is not dense in X.
   (d) f is bounded on some neighborhood of 0.

   [The proof that (a) implies (d) is clear. To show that (b) implies (c), if $N = \overline{N}$ and f is unbounded on $\{x : \|x\| < 1\}$ then $f(x/f(x)) = 1$. In particular, there is a point $x \in N$ with $\|x\| < 1$ such that $f(x) = 1 = f(x)$, and this leads to a contradiction.]

2. Show that any infinite dimensional normed linear space has a discontinuous linear functional. [Let $x_1, \ldots$ be an infinite sequence of linearly independent elements such that $\|x_n\| = 1$ for all n. Define f sequentially on the x_n and extend f to the whole space using linearity.]

3. Let $K(x, t)$ be continuous on $[0, 1] \times [0, 1]$. Show that $\|A\| = \max_{0 \le x \le 1} \int_0^1 |K(x, t)|\,dt$. [If $\int_0^1 |K(x_0, t)|\,dt$ is a maximum at the point x_0 and $K(x_0, t) = u(t)v(t)$ where $|u(t)| = 1$, let $v_1, v_2, \ldots$ be continuous functions such that $|v_n(t)| \le 1$ for all n and $\int_0^1 v_n(t)\,dt \to \int_0^1 |v(t)|\,dt$. Then $\|Av_n\| \le \|A\|$ and $\|Av_n\| = \int_0^1 K(x, t)v_n(t)\,dt$ by definition of A. Since $\int_0^1 |K(x_0, t)|\,dt \le \|A\|$ follows that $\int_0^1 |K(x_0, t)|\,dt \le \|A\|$.]

   The same argument, with integrals replaced by sums, shows that if A is a matrix operator on $\mathbb{C}^N$, with the norm $\|x\| = \max_{1 \le i \le N} \sum_{j=1}^N |a_{ij}|$.

4. Let M be a closed manifold in the normed linear space X. Define on X/M, the coset space modulo M, that is, $X/M = \{x + M : x \in X\}$, the function

$$\|\xi\| = \inf\{\|y\| : y \in \xi\}.$$

Show that $\|\xi\| = \inf\{\|x - m\| : m \in M\} = \operatorname{dist}(x, M)$. Show that the above formula indeed defines a seminorm on the quotient space X/M, a norm if M is closed.

5. If X is a Banach space and M is a closed subspace, show that X/M is a Banach space.

6. (a) Let A be a linear operator from X to M. Show that A is continuous with A^{-1} continuous on its domain $\mathcal{D}(A)$ iff there is a positive number $m > 0$ such that $\|Ax\| \geq m\|x\|$ for all $x \in \mathcal{D}(A)$.

(b) Let $\|\ \|_1$ and $\|\ \|_2$ be norms on the linear space X. Show that the norms induce the same topology (in other words, the open sets are the same for each norm) if there are positive numbers $m, M > 0$ such that

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1 \quad \text{for all} \quad x \in X.$$

[This may be done using part (a), or it may be shown directly that, for example, if $\|x\|_2 \leq \varepsilon$ then $\|x\|_1 \leq \varepsilon/m$ for all x (in the topology induced by $\|\ \|_1$ is easier than (that is, included in) the topology induced by $\|\ \|_2$.]

7. Let X be a finite-dimensional normed linear space, with dimension n, and let $\|\ \|_*$ be any norm on X, and define

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} \quad \text{where} \quad x = \sum_{i=1}^{n} x_i e_i.$$

(a) Show that for some positive real number k, $\|x\|_* \leq k\|x\|_2$ for all $x \in X$.

(b) Show that the unit sphere (real number) $\|x\|_2 \leq 1$ for $\|x\|_2 = 1$. By part (a), the map $x \to \|x\|_*$ is continuous in the topology induced by $\|\ \|_2$.

(c) Show that $\|x\|_* \geq m\|x\|_2$ for all $x \in X$, and conclude that all norms on a finite-dimensional space induce the same topology.

(d) Let M be an infinite-dimensional normed linear space, and let L be a finite-dimensional subspace of M. Show that L is closed in M.

8. (Riesz lemma) Let M be a closed proper subspace of the normed linear space L. Show that if $0 < \varepsilon < 1$, there is an element $x \in L$

such that $\|x_n\| = 1$ and $\|x - x_n\| \geq \varepsilon$ for all $n \in \mathbb{N}$. Choose $x_1 \neq 0$, and let $d = \dim\{x_1, Mx_1\} < 2$. Since M is closed, $\dim\{x_1, Mx_1\} < 2$ for some $x_2 \in \{x_1, Mx_1\}^\perp$ and set $v_2 = x_2 - \langle x_2, x_1\rangle x_1$ and $w_2 = v_2/\|v_2\|$.

9 Let X be a normed linear space. Show that the following are equivalent.

(a) X is finite-dimensional.

(b) X is topologically isomorphic to a Hilbert space $\mathcal{H}$ over $\mathbb{R}$ and X is a real vector space), that is, there is a one-to-one, onto continuous linear map between the two spaces.

(c) X is locally compact (every point in X has a neighborhood whose closure is compact).

(d) Every closed bounded subset of X is compact.

(e) The set $\{x \in X : \|x\| = 1\}$ is compact.

(f) The set $\{x \in X : \|x\| \leq 1\}$ is totally bounded, that is, can be covered by a finite number of open balls of any positive radius.

[Prove that (a) always implies (b), (b) implies (c), (c) implies (d), (d) implies (e), (e) implies (f), and (f) implies (a). To prove that (f) implies (a), use Problem 8.]

10 Let c be the vector space of convergent sequences of complex numbers (see Section 2.2(b)). If T is a continuous linear functional on c, show that there is a unique element $x = (x_0, x_1, \ldots) \in \ell^1$ such that for all $y \in c$

$$Ty = \left(\lim_{n\to\infty} y_n\right)\left[x_0 - \sum_{j=1}^{\infty} x_j\right] + \sum_{j=1}^{\infty} x_j y_j.$$

[Hint: Since $\|T\| = \sup_{\|y\|=1} |Ty| < \infty$, $\sum_{j=1}^{\infty} |x_j| < \infty$. Use this to show that the sequence of linear functionals converging to T has a representation of a continuous linear functional on ℓ^1 as in Problem 7: $f(x) = \sum_{j=0}^{\infty} x_j y_j$. Then x is unique and is determinable to T.]

11 Let $(X, \mathcal{M}, \mu)$ be a measure space.

(a) Assume μ is finite. Let $\mathcal{H} = L^2(X, \mathcal{M}, \mu)$. If T is a continuous linear functional on $\mathcal{H} = L^2(X, \mathcal{M}, \mu)$, show that there is an element $y \in \mathcal{H}$ such that

$$Tf = \int_X fy\,d\mu \quad \text{for all} \quad f \in L^2.$$

Furthermore, $\mu = |\nu_\mu|(E)$ for some $\nu_\mu \in L^1$. [Hint: Show that $\nu \ll \mu$ and use the Radon–Nikodým theorem.]

Exercise Theorem 3.3.1 remains true if $L^1(X, \mathcal{F}, \mu)$ and $L^\infty(X, \mathcal{F}, \mu)$ are replaced by $L^1(X, \mathcal{F}, \mu)$ and $L^\infty(X, \mathcal{F}, \mu)$, respectively, provided μ is σ-finite. [Hint: Let $\{X_n\}$ be a sequence of sets in $\mathcal{F}$ such that $X_n \subset X_{n+1}$, $\mu(X_n) < \infty$, and $X = \bigcup_{n=1}^{\infty} X_n$. Define

$$\nu_n(A) = \int_A \chi_{X_n} \, d\mu \qquad \text{for all } A \in \mathcal{F},$$

where χ_{X_n} is the characteristic function of X_n. Show that

$$\|\nu_n\| \leq \mu(X_n).$$

(i) If $g \in L^\infty(X, \mathcal{F}, \mu)$, show that $g = g\chi_{X_n}$ a.e. $[\mu]$ for some $n \in \mathbb{N}$. Thus $\int_X g \, d\nu$ may be obtained by passing to the limit in $\int_X g\chi_{X_n} \, d\nu$.

(ii) Let A_n, $n = 1, 2, \ldots$, be sets with $|\nu_n|(A_n) > 0$ and $\sup\{|\nu_n(B)| : B \subset A_n\} \leq \frac{1}{2}|\nu_n|(A_n)$. Show that ν_n converges in $L_\mathbb{R}^\infty$ to a finite functional ν, and ν is essentially independent of the particular sequence $\{X_n\}$. Furthermore, $\|\nu\| \leq \mu(X)$.

(iii) Show that $\nu(\chi_E) = \int_E g \, d\mu$ holds for all $E \in \mathcal{F}$. Since $|\nu|_n \uparrow |\nu|$ by (ii), $|\nu| \ll \mu$, by Hölder's inequality, we have $\|\nu\| = \|g\|_q$. Thus the result of (i) holds for arbitrary g.

(iv) Prove (i) with $p = 1$, $q = \infty$.

(v) Prove (i) with $p = 1$ and $q = \infty$. [For an extension of this result, see Kelley and Namioka (1963, Problem 14.N).]

It follows that there is an isometric isomorphism of L^{p*} and L^q if $1 < p < \infty$, $p^{-1} + q^{-1} = 1$, and μ is σ-finite. The theorem remains valid for $p = 1$, $q = \infty$.

3.4 Basic Theorems of Functional Analysis

Almost every result of functional analysis leans heavily on at least one of the three basic principles of this section: the Hahn–Banach theorem, the principle of uniform boundedness, and the open mapping theorem. We are going to establish these results and discuss applications.

We have considered an expression such as $\int_0^1 f(t)\,dt$ as a linear functional defined on a subspace M of a vector space L. There is no difficulty in extending f to a linear functional on all of L: simply extend a Hamel basis of M to a Hamel basis for L, define f arbitrarily on the basis vectors not belonging to

M, and scored by linearity. However, $\mathcal{L}$ is normed and we desire that the extension of f have the same norm as the original functional. The problem has been solved in Theorem 1. We first prove a preliminary result.

1.6.1 Lemma. Let L be a real vector space and M a subspace of L, and let p be a map from L to $\mathbb{R}$ satisfying

$$p(x+y) \le p(x) + p(y) \qquad \text{for all } x, y \in L$$

$$p(\alpha x) = \alpha p(x) \qquad \text{for all } x \in L \text{ and all } \alpha > 0.$$

The first property is called subadditivity, the second positive-homogeneity. Note that positive-homogeneity implies that $p(0) = 0$ (let $x = 0$ in certain $p(\alpha) = \alpha p(x)$ for all $\alpha > 0$). A subadditive, positive-homogeneous map is sometimes called a *sublinear* functional.

Let M be a subspace of L and g a linear functional defined on M such that $g(x) \le p(x)$ for all $x \in M$. Let x_0 be a fixed element of L. For any real number α, the following are equivalent:

(i) $g(x) + \alpha x_0 \le p(x + \alpha x_0)$ for all $x \in M$ and all $\lambda \in \mathbb{R}$.

(ii) $-p(-x - \alpha x_0) - g(x) \le \alpha \le p(x + \alpha x_0) - g(x)$ for all $x \in M$.

Furthermore, there is a real number α satisfying (ii), and hence (i).

Proof. To prove that (i) implies (ii), first set $\lambda = 1$, and then set $\lambda = -1$ and replace x by $-x$. This yields $g(x) - \alpha \le p(x - \alpha x_0)$ by linearity. Conversely, if (ii) holds and $\lambda > 0$, replace x by x/λ in the right-hand inequality of (ii) and multiply by λ; replace x by x/λ in the left-hand inequality. In either case, the positive homogeneity of p yields (i). If $\lambda = 0$, (i) is true by hypothesis.

To produce the desired α, let x and y be arbitrary elements of M. Then

$$g(x) - g(y) = g(x - y) \le p(x - y) \qquad (x \in M, -y \in M)$$

$$= p(x + x_0 + y - x_0) \le p(x + x_0) + p(y - x_0) \qquad \text{by subadditivity.}$$

It follows that

$$\sup_{y \in M} [-p(y - x_0) + g(y)] \le \inf_{x \in M} [p(x + x_0) - g(x)].$$

Any α between the sup and the inf will work.

We may now prove the main extension theorem.

1.6.2 Hahn-Banach Theorem. Let p be a subadditive, positive-homogeneous functional on the real linear space L, and g a linear functional on the

subspace M, with $g \leq p$ on M. Prove that g has a linear functional f on L such that $f = g$ on M and $f \leq p$ on all of L.

PROOF. If $M = L$, there is nothing to prove. If $M \neq L$, choose $x_1 \in L$ with $x_1 \notin M$. Let M_1 consist of all elements $x + \alpha x_1$, where $x \in M$ and α is any real number. We shall show that g can be extended to a linear functional g_1 on M_1 such that $g_1 \leq p$ on M_1.

Now g_1 has the prescribed value at the point x_1 if and only if, for every $x \in M$ and every positive α,

$$g(x) - \alpha g(x_1) \leq p(x - \alpha x_1) \quad \text{and} \quad g(x) + \alpha g(x_1) \leq p(x + \alpha x_1).$$

Dividing by α and setting $x = \alpha y$, we see that g_1 has the prescribed value at x_1 if and only if

$$g(y) - g(x_1) \leq p(y - x_1) \quad \text{and} \quad g(y) + g(x_1) \leq p(y + x_1)$$

for all $y \in M$. Thus we must choose $g(x_1)$ so that

$$g(y) - p(y - x_1) \leq g(x_1) \leq p(y + x_1) - g(y) \quad \text{for all } y \in M.$$

Such a choice is possible because $g(y_1) - p(y_1 - x_1) \leq p(y_2 + x_1) - g(y_2)$ for all $y_1, y_2 \in M$, since

$$g(y_1) + g(y_2) = g(y_1 + y_2) \leq p(y_1 + y_2) \leq p(y_1 - x_1) + p(y_2 + x_1).$$

Define $g_1(x_1)$ to be any real number c such that

$$\sup_{y \in M} [g(y) - p(y - x_1)] \leq c \leq \inf_{y \in M} [p(y + x_1) - g(y)].$$

Then $g_1(x + \alpha x_1) = g(x) + \alpha c$ for all $x \in M$ and all real α. Define $P(x) = p(x + x_0) - g(x_0)$. Then $g_1(x) \leq P(x)$ for all $x \in M_1$. It follows that g_1 is a linear functional on M_1 and $g = g_1$ on M. For g_1 is additive since it is g on M. If $a, b \in \mathbf{R}$ we have

$$g_1(ax) = g_1(ax) + g_1(bx) = g_1(ax + bx) = g_1(x)$$

$$= ag_1(x) + bg_1(x) = ag_1(x) + bg_1(x).$$

$$= \frac{1}{2}[p(x+iy) + p(-x-iy)],$$
$$= \frac{1}{2}[p(x+iy) + p(x+iy)].$$

Note that homogeneity of f does not immediately imply homogeneity of f for $f(\lambda x) = \lambda f(x)$ for real λ, but not in general for complex λ. [For example, let $f = \mathbb{C}$, $f(x+iy) = x - iy$, $\lambda = i$.]

We now prove the complex version of the Hahn–Banach theorem.

3.4.3 Theorem. Let L be a vector space over $\mathbb{C}$, and p a seminorm on L. If g is a linear functional on the subspace M and $|g(x)| \le p(x)$ on M, there is a linear functional f on L such that $f = g$ on M and $|f(x)| \le p(x)$ on L.

Proof. Since p is a seminorm, it is subadditive and absolutely homogeneous over the real scalars. If $g_1 = \operatorname{Re} g$, then $g(x) = g_1(x) - ig_1(ix)$, by (3.4.2). There is an extension f_1 of g_1 to a linear map f_1 of L to $\mathbb{R}$ such that $f_1 = g_1$ on M and $|f_1| \le p$ on L. Define $f(x) = f_1(x) - if_1(ix)$. Then f is a complex linear functional on L and $f = g$ on M. For $x \in L$ and let $f(x) = re^{i\theta}$, $r \ge 0$. Then

$$\begin{aligned} |f(x)| &= r = e^{-i\theta} f(x) \\ &= f(e^{-i\theta}x) && \text{since } r \text{ is real} \\ &\le p(e^{-i\theta}x) && \text{since } f_1 \le p \text{ on } L \\ &= p(x) && \text{by absolute homogeneity.} \end{aligned}$$

3.4.4 Corollary. Let g be a continuous linear functional on the subspace M of the normed linear space L. There is an extension of g to a continuous linear functional f on L such that $\|f\| = \|g\|$.

Proof. Let $p(x) = \|g\| \cdot \|x\|$; then p is a seminorm on L and $|g| \le p$ on M by definition of $\|g\|$. The result follows from (3.4.3).

A typical application of the Hahn–Banach theorem is the result that in a normed linear space, there are enough continuous linear functionals to distinguish points. In other words, if $x \neq y$, there is a continuous linear functional f such that $f(x) \neq f(y)$. We now prove this along with a more general result.

3.4.5 Theorem. Let M be a subspace of the normed linear space L, and let $M^\perp$ be the collection of all continuous linear functionals on L.

(a) If $x_0 \notin M$, there is an $f \in L'$ such that $f = 0$ on M, $f(x_0) = 1$, and $\|f\| = 1/d$, where d is the distance from x_0 to M.

(b) $x_0 \in M$ iff each $f \in M^\perp$ that vanishes on M also vanishes at x_0.

for $x \neq 0$ there is an $f \in L'$ such that $\|f\| = 1$ and $f(x) = \|x\|$; thus the maximum value of $|f(x)|/\|x\|$, $x \neq 0$, is achieved at x. In particular, if $x \neq y$, there is an $f \in L'$ such that $f(x) \neq f(y)$.

PROOF (a) First note that $\{f(x) : \|x\| \leq 1\}$ is the set of all elements $\alpha = \alpha_1 + \alpha_2 i$ with $|\alpha| \leq 1$, and since $\eta \notin$ the set is uniquely determined by η. Define f on $M = \{\beta x : \beta \in \mathbb{C}\}$ by $f(\beta x) = \beta\eta$; f is linear, and furthermore, $\|f\| = 1$, as we now prove. By (3.3.1) we have

$$\|f\| = \sup\left\{\frac{|f(z)|}{\|z\|} : z \in M,\ z \neq 0\right\} = \sup\left\{\frac{|\alpha|}{|\alpha - \alpha_0|} : \alpha \in \mathbb{C},\ |\alpha| \neq 1 \text{ or } \alpha = 0\right\} = \sup\left\{\frac{\alpha}{\alpha - \alpha_0} : \alpha \in \mathbb{C},\ |\alpha| \neq 1\right\}$$

since $f(z) = 0$ when $\alpha = 0$. Now

$$\left|\frac{\alpha}{\alpha - \alpha_0}\right| = \left|1 + \frac{\alpha_0}{\alpha - \alpha_0}\right| \geq 1 - \frac{|\alpha_0|}{|\alpha - \alpha_0|} \quad \text{for some } \alpha \in \mathbb{C},$$

hence $\|f\| = \sup\{|\alpha/(\alpha - \alpha_0)| : |\alpha| \neq 1\} = 1$ (why?). The result now follows from (3.4.2).

(b) This is immediate from (a).

(c) Apply (a) with $M = \{0\}$. In order to find $g \in L'$ with $g(x) = \|x\|$ and $\|g\| = 1$, set $g(\beta x) = \beta\|x\|$ and $g(y) = 0$ for $y \in M$. ∎

The Hahn–Banach theorem is certainly not trivial, but the concept of reflexivity, which we now discuss, is. If L is a normed linear space, and L' the set of continuous linear functionals on L, L' is a normed space, so that we may also consider L'', the conjugate space of L', or the second conjugate space of L. We may similarly define L''', and so on. It is a subspace of L'' as follows. If $x \in L$, we define $x'' \in L''$ by

$$x''(f) = f(x), \quad f \in L'.$$

If $\|f_n - f\| \to 0$, then $f_n(x) \to f(x)$, hence x'' is in fact a continuous linear functional on L'. Let us examine the mapping $x \to x''$ of L into L''.

3.4.4 *Theorem.* Define $A : L \to L''$ by $Ax = x''$. Then A is an isometric isomorphism of L onto the subspace $\{x'' : x \in L\}$ of L''; therefore, if $x \in L$, we have by (3.3.1), $\|x\| = \|x''\| = \sup\{|f(x)| : f \in L',\ \|f\| \leq 1\}$.

Proof. To show that π is linear, we write

$$\pi(\alpha x + \beta y)(f) = f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) = \alpha \pi(x)(f) + \beta \pi(y)(f).$$

We now prove that π is an isometry: $\|\pi(x)\| = \|x\|$ for all $x \in M$. Since consequently, it is only to show that $x \in M \Rightarrow \|\pi(x)\| \geq \|x\|$. This is trivial if $x = 0$, and hence $\|\pi(x)\| \geq \|x\|$. On the other hand, by 3.4.3(c), there is an $f \in M^*$ such that $\|f\| = 1$ and $f(x) = \|x\|$. Thus

$$\sup\{|\pi(x)(f)| : f \in M^*, \|f\| = 1\} \geq \|x\|,$$

so that $\|\pi(x)\| \geq \|x\|$, and consequently $\|\pi(x)\| = \|x\|$.

If $\pi(X) = X^{**}$, X is said to be *reflexive*. Note that X^{**} is complete by 3.1.5 and so, by 3.4.6, a reflexive normed linear space is necessarily complete. We shall now consider some examples.

3.4.7 *Examples.* (a) Every Hilbert space is reflexive. For if H is the conjugate isometry of $H^* \to \mathcal{B}(H, \mathbb{F})$ becomes a $\mathbb{F}$-bilinear form on $H \times H$ such that $\langle f, g \rangle = \overline{\langle g, f \rangle}$. Thus if $g \in H^{**}$, we have, for some $x \in H$, $g(f) = \langle x, f \rangle = \overline{\langle f, x \rangle} = f(x)$, where $x = \delta(g)$. Therefore $g = \pi(x)$.

(b) If X is reflexive, X^* is reflexive. For by 3.4.6(b), $\pi(X)$ is weak-* dense in X^{**}, where $\pi(x)(f) = f(x)$. This shows that $\pi(X)$ is an algebraic complement of $\pi(X)^\perp = X^*$. But then $X = \delta(g)$.

(c) Essentially the same argument, with the aid of Problem 4 of Section 3.1, shows that if $(\mathcal{S}, \mu)$ is an arbitrary measure space, $L^1(\mathcal{S}, \mu)$ is reflexive if $L^\infty(\mathcal{S}, \mu)$ is reflexive.

(d) The space l^1 is not reflexive. This is a consequence of the following result.

3.4.8 *Theorem.* If X is a normed linear space and X^* is separable, so is X. Thus if X is reflexive and separable (or more generally, if X^* is separable), then so is X^*.

Proof. Let $\{f_1, f_2, \ldots\}$ be a countable dense subset of $\{f \in X^* : \|f\| = 1\}$. Recall that any subset of a separable metric space is separable. Since $\|f_n\| = 1$, we can find points $x_n \in X$ with $\|x_n\| = 1$ and $f_n(x_n) \geq \frac{1}{2}$ for all n. Let M be the space spanned by the x_n; we claim that $M = X$. If not, 3.4.3(b) yields an $f \in X^*$ with $f \neq 0$ on M and $\|f\| = 1$. But then $\|f - f_n\| \geq |f(x_n) - f_n(x_n)| = |f(x_n)| \geq \frac{1}{2}\|f\|$ for all n, contradicting the assumption that $\{f_1, f_2, \ldots\}$ is dense. $\square$

To establish the second assertion, we note that X is separable since $\pi : X \to \pi(X)$ is an isometry and $\pi(X)$ is dense. If $x \in X$ and $x^{(n)} = (x_1, \ldots, x_n, 0, 0, \ldots)$, then

ℓ^∞ ... ℓ^∞. But ℓ^1 ... by 2.5.4(c), and this space is not separable. For ... $\|e_j - e_k\| = 2^{1/p}$ ... $\neq 1$ for all j, there is no countable subset ... $\|x - e_j\| < 1$ ... x. Thus the sets $O_j = \{x \in \ell^\infty : \|x - e_j\| < \frac{1}{2}\}$, $j \in \mathbb{N}$, form an uncountable family of disjoint open sets. If there were a countable dense ... then each O_j would have to contain ... in it, which is impossible. Thus ℓ^∞ is not separable.

We now consider the second basic result of this section. Suppose that the ... belonging to the ordinary unit ball S are bounded linear operators from X to W, where X and W are normed linear spaces. The uniform boundedness principle asserts that if X is complete and ... $\sup\{\|A_\alpha x\| : \alpha \in \Lambda\}$ is finite for each $x \in X$, then the A_α are uniformly bounded, that is, $\sup\{\|A_\alpha\| : \alpha \in \Lambda\}$ is finite. Completeness of X is necessary, as we see from ... Let Σ be the set of all sequences $x = (\xi_1, \xi_2, \ldots)$ of complex numbers such that $\xi_n = 0$ for all but finitely many n, with the l^2 norm. Let $\mathbb{N} = \mathbb{P}$, and $A_n x = n\xi_n$, $n = 1, 2, \ldots$. Just as in ... for each $x \in \Sigma$, so the A_n are pointwise bounded, although $\|A_n\| = n \to \infty$.

The proof that you shall give uses the Baire category theorem, which states that if a complete metric space is a countable union of closed sets, one of the sets must have a nonempty interior.

3.4.9 Principle of Uniform Boundedness. Let $\{A_\alpha : \alpha \in \Lambda\}$ be a family of bounded linear operators from a Banach space X into a normed linear space W. If the A_α are pointwise bounded, they are uniformly bounded.

Proof. For $n = 1, 2, \ldots$, let $C_n = \{x \in X : \sup_\alpha \|A_\alpha x\| \leq n\}$. Since the A_α are pointwise bounded, $\bigcup_{n=1}^\infty C_n = X$, and since the A_α are continuous, each C_n is closed. By the Baire category theorem, for some n there is a closed ball $S' = \{x \in X : \|x - x_0\| \leq r\} \subset C_n$. Now if $\|y\| \leq 1$ and $x \in S'$, we have

$$\|A_\alpha y\| = \frac{2}{r}\|A_\alpha(x_0 + \tfrac{r}{2}y) - A_\alpha(x_0 - \tfrac{r}{2}y)\| \leq \frac{2}{r}\|A_\alpha(x_0 + \tfrac{r}{2}y)\| + \frac{2}{r}\|A_\alpha(x_0 - \tfrac{r}{2}y)\| \leq \frac{2n}{r} + \frac{2n}{r} \quad \text{since } x_0 + \tfrac{r}{2}y \text{ and } x_0 - \tfrac{r}{2}y \text{ belong to } S'.$$

Thus $\|A_\alpha\| \leq 4n/r$. ∎

The uniform boundedness principle is used in an important way in the study of weak convergence, a concept that we may describe

5.1.10 Definitions and Comments. The sequence $\{x_n\}$ in the normed linear space E is said to converge weakly to $x \in E$ iff $f(x_n) \to f(x)$ for every $f \in E'$. [Notation: $x_n \xrightarrow{w} x$.] Convergence in the metric of E ($\|x_n - x\| \to 0$) will be called strong convergence. We will be very careful to say $x_n \to x$ [in this terminology, pointwise convergence of the sequence of linear operators A_n to the linear operator A means strong convergence of $A_n x$ to Ax for each x (see 3.3.6).]

It follows from the preceding that strong convergence implies weak convergence. To see that the converse does not hold, let $\{e_1, e_2, \ldots\}$ be an infinite orthonormal sequence in a Hilbert space H. Let $x \in H$. Then $(x, e_n) \to 0$ as $n \to \infty$ by Bessel's inequality (lemma 4.6.3). But $\|e_n\| = 1$, so e_n does not converge strongly to 0.

In a finite-dimensional space, strong and weak convergence coincide. For in $\mathbb{R}^n$ or $\mathbb{C}^n$, $x_n = (\xi_{n1}, \ldots, \xi_{nn})$ converges weakly to $x = (\xi_1, \ldots, \xi_n)$ if and only if $\xi_{nj} \to \xi_j$ for $j = 1, \ldots, n$, since the continuous linear functionals on $\mathbb{R}^n$ are the coordinate functions $f(x) = \xi_j$. Then $f(x_n) \to f(x)$ so that $\xi_{nj} \to \xi_j$ as $n \to \infty$, $j = 1, \ldots, n$. Thus $x_n \to x$ strongly.

We give a few properties of weak convergence.

5.1.11 Theorem. (a) A weakly convergent sequence $\{x_n\}$ is bounded, that is, $\sup_n \|x_n\| < \infty$. In fact if $x_n \xrightarrow{w} x_0$, then $\|x_0\| \le \liminf_{n\to\infty} \|x_n\|$.

(b) If A is a bounded linear operator from E to H and the sequence $\{x_n\}$ converges weakly to x_0 in E, then Ax_n converges weakly to Ax_0 in H.

(c) If the sequence $\{x_n\}$ converges weakly to x_0, then any subsequence has the subsequence spanned by the x_n.

(d) If M is a linear manifold in E, then M is the weak closure of some sequence in M. Then x is the strong limit of some sequence in M.

Proof. (a) The x_n may be regarded as continuous linear functionals on E' with $x_n(f) = f(x_n)$, $f \in E'$. The x_n are pointwise bounded on E' since $f(x_n) \to f(x_0)$. Hence by the uniform boundedness principle, $\sup_n \|x_n\| = \sup_n \sup\{|f(x_n)| : f \in E', \|f\| \le 1\} < \infty$. Also, if $f \in E'$,

$$|f(x_0)| = \lim_{n\to\infty} |f(x_n)| = \liminf_{n\to\infty} |f(x_n)| \le \|f\| \liminf_{n\to\infty} \|x_n\|.$$

Since $\|x_0\| = \sup\{|f(x_0)| : f \in E', \|f\| \le 1\}$, we must conclude that $\|x_0\| \le \liminf_{n\to\infty} \|x_n\|$.

(b) If $g \in H'$, define $f = g \circ A$. f is A as continuous. We have $f \in E'$, so that $f(x_n) \to f(x_0)$, that is, $g(Ax_n) \to g(Ax_0)$. But g is arbitrary, hence $Ax_n \xrightarrow{w} Ax_0$.

(a) If this is false, 3.4.5(a) yields an $f \in L^p$ with $f(x_n) = 1$ and $f(y) = 0$, contradicting $x_n \xrightarrow{w} y$.

(b) Let $\{x_n\}$ be a sequence in M converging weakly to x. By (a), x is a convex combination of finitely many of the x_n. However, these finite linear combinations belong to M since M is a subspace. Thus x is the strong limit of a sequence in M. ∎

In order to characterize weak convergence in Hilbert spaces, the following result is useful.

3.4.12 Theorem. Let E be a subset of H such that $\overline{\mathrm{sp}}(E) = H$, and let $\{x_n\}$ be a sequence in H. If $x_n \to x$, then $x_n \xrightarrow{w} x$ if and only if $\sup_n \|x_n\| < \infty$ and $f(x_n) \to f(x)$ for all $f \in E$.

Proof. The "only if" part follows from 3.4.11 and the definition of weak convergence. For the "if" part, let $f \in H'$, and choose elements $f_k \in \mathrm{sp}(E)$ with $\|f - f_k\| \to 0$. Then

$$|f(x_n) - f(x)| \le |f(x_n) - f_k(x_n)| + |f_k(x_n) - f_k(x)| + |f_k(x) - f(x)|$$
$$\le \|f - f_k\| \|x_n\| + |f_k(x_n) - f_k(x)| + \|f_k - f\| \|x\|.$$

Since $\{x_n\}$ is bounded in norm, given $\varepsilon > 0$, we may choose k such that the right-hand side is at most $\|f - f_k\|(\|x_n\| + \|x\|) + \varepsilon$ and $f_k(x_n) \to f_k(x)$ as $n \to \infty$ since $f_k \in \mathrm{sp}(E)$; and since ε is arbitrary, the result follows. ∎

We now describe weak convergence in L^p and l^p.

3.4.13 Theorem. Assume $1 < p < \infty$.

(a) Let $x_n = (\xi_{n1}, \xi_{n2}, \ldots) \in l^p$, $n = 1, 2, \ldots$. If $y \in l^p$, then $x_n \xrightarrow{w} y$ if $\sup_n \|x_n\| < \infty$ and $\xi_{nk} \to \eta_k$ as $n \to \infty$ for each k.

(b) Let μ be σ-finite on $(S, \mathcal{F})$, $1 < p < \infty$, where q is assumed finite. If $g \in L^p(S, \mathcal{F}, \mu)$, then $g_n \xrightarrow{w} g$ if and only if $\sup_n \|g_n\| < \infty$ and $\int_A g_n \, d\mu \to \int_A g \, d\mu$ for each $A \in \mathcal{F}$. (It will also be convenient to allow the dimension between L^p and L^q and treat the elements of L^p as functions rather than equivalence classes.)

Proof. (a) Define $f_k \in (l^p)'$ by $f_k(x) = \xi_k$. Then f_k corresponds to the sequence in l^q with a one in position k and zeros elsewhere [see 3.4.4(c)]. Take $E = \{f_1, f_2, \ldots\}$ and apply 3.4.12.

(b) For each $n = 0, 1, \ldots$, define $f_n \in (L^p)'$ by $f_n(x) = \int_S g_n x \, d\mu$. f_n corresponds to the indicator function $\chi_A \in L^q$ (see Problem 11, Section 3.3). Take

Since the set of all p-th roots of f is finite, f is locally H^∞ because the simple functions are dense in H^∞, and the result follows from 3.4.12. $\square$

Note that Theorem 3.4.13 holds also when $p = 1$ since analytic functions are dense in T^p.

We now consider the Banach base problem for open mapping theorems. This will allow us to conclude that under certain conditions the inverse of a one to one continuous linear transformation is continuous. We cannot make this assertion in general as the following example shows. Let T be the set of all continuous complex-valued functions f on $[0, 1]$ such that $f(0) = 0$, $M = \|f\| + \|f'\|$ is finite and f' is continuous on $[0, 1]$. Let the sup norm on L and M, L and M defined by $\|f\| = \sup_{0 \le t \le 1} |f(t)|$. Then A is one-to-one bounded linear operator from L onto M. The problem with A^{-1} is now in showing that it is not.

But A^{-1} is discontinuous. For example if $x_n(t) = \sin nt$, then $\|x_n(t)\| = \max_{0 \le t \le 1} |x_n(t)| = 1$ — namely $x_n = \frac{1}{n}$ in M, but x_n is not in T. A hypothesis under which continuity of the inverse holds is one which connects the two spaces L and M. In the above example, M is not complete. For example, a sequence of polynomials may converge uniformly to a continuous function without even having derivatives. In fact it can converge to a discontinuous function.

We now state the third basic theorem.

3.4.14 Open Mapping Theorem Let A be a one to one linear transformation from the Banach space L onto the Banach space M. Then A is an open map, that is if D is an open subset of L, then $A(D)$ is an open subset of M. Consequently if A is also continuous, then A^{-1} is bounded.

Proof Let $B_r = B(0, r)$ be the open ball with center at 0 and radius r in L. It suffices to show that $A(B_r)$ contains a ball with center at 0 in M, since translations of neighborhoods about arbitrary points are translations of neighborhoods about 0. Now $L = \bigcup_{n=1}^{\infty} B_n$ and A maps onto M, so $M = \bigcup_{n=1}^{\infty} A(B_n)$. Since M is complete, we can conclude from the Baire category theorem that one of the $\overline{A(B_n)}$ is not nowhere dense. Since $A(rx) = rA(x)$, it follows that $\overline{A(B_r)}$ is not nowhere dense. Thus for some $y_0 \in M$ and $r > 0$, the ball $B(y_0, 4r)$ is contained in the closure of $A(B_1)$. It follows that we may select $x_1 = A x_1$ in $A(B_1)$ such that $\|y_0 - y_1\| < 2r$. If y_1 is already very close to the boundary of $A(B_1)$, we choose y_1 by a very slight variation as $A(\frac{1}{2}x_1)$; this element can be chosen so that it is closer to y_0 is $\|y_0 - A x_1\|$. By the triangle inequality,

$$B(y_1, 2r) \subset B(y_0, 4r) \subset \overline{A(B_1)}. \tag{1}$$

We claim that

$$\text{if } \|y - y_1\| < 2r \text{ then } y \in A(B_1). \tag{2}$$

To see this, note that $x = \lambda x_1 + (1-\lambda)x_2$, and since $y_1 = Ax_1$ the second factor is $\lambda A(B_1)$ by (1). Therefore x is in the plus the limit of a sequence in $A(B_1)$, so that $y \in \overline{A(B_1)} = W_1$ (but $y_1 = Bx_1$, so $x_1 \in A^{-1}W_1$, and consequently $x_1 \in B_1$, because (2)).

Again because the B_i differ only by a scalar factor, we can repeat the above argument with B replaced by B_i for any $i > 0$. The inclusions (1) and (2) become:

$$\text{if } \|x\| < \delta r \quad \text{then} \quad x \in A(B_1). \tag{3'}$$

We are going to show that if $\|y\| < r\delta$, then $y \in A(B_1)$. Since $A(B_1) \subset W_1$, and hence by scaling the image of any ball with center at 0 in X contains a ball with center at 0 in W, completing the proof. We use (3') to generate an inductive procedure:

Step 1 — $i = 2$: choose $x_1 \in B_1$ such that $\|y - Ax_1\| < r\delta$.

Now apply (3') with $i = 2$: Choose $x_2 \in B_2$ such that $\|y - Ax_1 - Ax_2\| < r\delta^2$.

Now apply (3') once $i = 1, 0$, choose $x_3 \in B_3$ so that $\|y - Ax_1 - Ax_2 - Ax_3\| < r\delta^3$ and continue in this fashion. In general, we select $x_i \in B_i$ so that

$$\|y - \sum_{j=1}^{i} Ax_j\| < r\delta^i. \tag{4'}$$

If $s_n = \sum_{j=1}^{n} x_j$, then for $n > m$ we have $\|s_n - s_m\| \le \sum_{j=m+1}^{n} \|x_j\| < \delta^m$ since $\|x_j\| < \delta^j$. The conclusion of (3') implies that $\sum_{j=1}^{\infty} x_j$ converges. If x is the sum of the series then $\|x\| \le \sum_{j=1}^{\infty} \|x_j\| < \sum_{j=1}^{\infty} \delta^j = 1$, and by (4'), $Ax = y$. Therefore $y \in A(B_1)$. □

The open mapping theorem allows us to prove the closed graph theorem, which is often useful in proving that a particular linear operator is bounded.

First we observe that if X and M are normed linear spaces, we may define a norm on the product space $L \times M$ by $\|(x, y)\| = (\|x\|^p + \|y\|^p)^{1/p}$, $x \in L$, $y \in M$, where p is any fixed real number in $[1, \infty)$. For if $x, z \in L$, $y, y' \in M$,

$$\begin{aligned}(\|x - z\|^p + \|y + y'\|^p)^{1/p} &\le (\|x\| + \|z\|)^p + (\|y\| + \|y'\|)^p)^{1/p} \\ &\le (\|x\|^p + \|y\|^p)^{1/p} + (\|z\|^p + \|y'\|^p)^{1/p} \\ &\quad \text{by Minkowski's inequality applied to } \mathbb{R}^2\end{aligned}$$

Therefore the triangle inequality is satisfied and we have defined a norm on $L \times M$. (Other choices of p also work, but are unsatisfactory.) Furthermore, $(x_n, y_n) \to (x, y)$ if $x_n \to x$ and $y_n \to y$ (this magnitude of the vector p is not

norm on $X \times Y$ induces the product topology. Also, if X and Y are complete, so is $X \times Y$. [The same result is obtained using the analog of the l^1 norm, that is, $\|(x, y)\| = \max\{\|x\|, \|y\|\}$.]

5.4.25 Definition. Let A be a linear operator from X to Y, where X and Y are normed linear spaces. We say that A is closed if the graph $G(A) = \{(x, Ax) : x \in \mathcal{D}\}$ is a closed subset of $X \times Y$. Equivalently, A is closed if the following condition holds:

If $x_n \in \mathcal{D}$, $x_n \to x$, and $Ax_n \to y$, then $x \in \mathcal{D}(A)$ and $Ax = y$; in other words, $y = Ax$. This formulation shows that every bounded linear operator is closed. The converse holds if X and Y are Banach spaces.

5.4.26 Closed Graph Theorem. *If A is a closed linear operator from the Banach space X to the Banach space Y, then A is bounded.*

Proof. Since $G(A)$ is a closed subspace of $X \times Y$, it is a Banach space. Define $P: G(A) \to X$ by $P(x, Ax) = x$. Then P is linear and maps onto X, and $\|P(x, Ax)\| = \|x\| \leq \|(x, Ax)\|$; hence $\|P\| \leq 1$ so that P is bounded. Moreover, if $(x_n, Ax_n) \to (x, y)$, then $x_n \to x$, since convergence in $X \times Y$ implies convergence in X. Similarly, the linear operator $Q: G(A) \to Y$ given by $Q(x, Ax) = Ax$ is bounded. If $P(x, Ax) = 0$, then $x = 0$ and $Q(x, Ax) = 0$, so P is one to one. By 5.3.11, P^{-1} is bounded, and since $Q = QP^{-1}$, A is bounded. ■

As an application of the closed graph theorem, we show that if M is the projection of a Banach space X on a closed subspace M, then P is continuous (see 3.2.3(c)). Let x_n be a sequence of points in X with $x_n \to x$ and $Px_n \to y$. We must show that y is the element Px. Recall that in defining a projection operator we assume that X is the direct sum of the closed subspaces M and N; thus

$$x_n = x_n + z_n \quad \text{where} \quad x_n = Px_n \in M, \quad z_n \in N.$$

Since $x_n \to x$ and $x_n \to y$, it follows that $z_n = x_n - x_n \to x - y$. Since N is closed, $x - y = z \in N$. Also, $y \in M$, since M is closed. Finally, $x = y + z$, so $Px = y$. By 5.4.26, P is continuous.

Problems

1. Show that a subadditive, absolutely homogeneous functional on a vector space must be nonnegative, and hence a seminorm. Give an example of a subadditive, positive-homogeneous functional that fails to be nonnegative.
2. Let $(X, \mathcal{F}, \mu)$ be a measure space, and assume that $\mathcal{F}$ is countably generated, that is, there is a countable set $\mathcal{E} \subset \mathcal{F}$ such that $\sigma(\mathcal{E}) = \mathcal{F}$. Show that

maximal field $\mathscr{M}_{\mathscr{A}}$ over $\mathscr{A}$ is also commutative. See Problem 9, 3.3.) If μ is nonnegative on $\mathscr{F}_0$ and σ-subadditive, show that $\mu^*(A) = \mu(A)$ for $A \in \mathscr{F}_0$. If μ is arbitrary and $\mathscr{F}_0$ contains an infinite collection of disjoint sets $A_n \in \mathscr{F}_0$ with $\mu(A_n) > \varepsilon > 0$, show that $\mathrm{Dom}(\mu^*) = \mathscr{M}_{\mu^*}$ is not reflexive.

3. If L and M are normed linear spaces and $[L, M]$ is complete, show that M must be complete.

4. Let $A \in [L, M]$, where L and M are normed linear spaces. The adjoint of A is an operator $A^* : M^* \to L^*$ defined as follows: If $f \in M^*$ we take $(A^* f)(x) = f(Ax)$, $x \in L$. Establish the following results:
   (a) $\|A^*\| = \|A\|$.
   (b) $(\alpha A + \beta B)^* = \alpha A^* + \beta B^*$ for all $\alpha, \beta \in \mathbb{C}$ if $A, B \in [L, M]$.
   (c) If $A \in [L, M]$, $B \in [M, N]$, then $(BA)^* = A^* B^*$, where BA is the composition of A and B.
   (d) If $A \in [L, M]$, A has an inverse in M and A^* has an inverse in $[M^*, L^*]$, then $(A^*)^{-1} = (A^{-1})^*$.

5. Define the annihilator of the subset S of the normed linear space L as $S^\perp = \{f \in L^* : f(x) = 0 \text{ for all } x \in S\}$. Similarly, if $T \subset L^*$, define $T_\perp = \{x \in L : f(x) = 0 \text{ for all } f \in T\}$. If $A \in [L, M]$, we denote by $\mathscr{N}(A)$ the null space $\{x \in L : Ax = 0\}$, and by $\mathscr{R}(A)$ the range of A, that is, the closure of $\{Ax : x \in L\}$. Establish the following:
   (a) For any $S \subset L$, $S^{\perp\perp} = \overline{\mathscr{M}(S)}$, the closure spanned by S.
   (b) $\mathscr{N}(A^*) = \mathscr{R}(A)^\perp$ and $\overline{\mathscr{R}(A)} = \mathscr{N}(A^*)_\perp$.
   (c) $\overline{\mathscr{R}(A^*)} = M^*$ if and only if A is one-to-one.
   (d) $\mathscr{N}(A^*) = \mathscr{R}(A)^\perp$.
   (e) For any $A \in [L, M]$, $\mathscr{N}(A) = \mathscr{R}(A^*)^\perp$, $\mathscr{N}(A^*) = \mathscr{R}(A)^\perp$ if L is reflexive.
   (f) $\mathscr{N}(A) = \mathscr{N}(A^*A)$; $\mathscr{R}(A^*) = \mathscr{R}(A^*A) = \mathscr{R}(A^*A)^{\perp\perp}$ if L is reflexive.
   (g) If $\mathscr{R}(A) = M$, then A is invertible—that is, the inverse exists if L is reflexive.

6. Consider the Hahn–Banach theorem 5.4.3, with the additional assumption that L is a normed linear space (or more generally, a topological vector space) and φ is continuous on $\mathscr{M}$, hence continuous on L since $|p(x) - p(y)| \le p(x - y)$. Show that if L is separable, the theorem may be proved without Zorn's lemma. It follows that 5.4.3 and 5.4.4 do not require Zorn's lemma under the above hypotheses.

7. If μ_0 is a finitely additive, nonnegative real valued set function on a field $\mathscr{F}_0$ of subsets of a set Ω, use the Hahn–Banach theorem to show that μ_0 has an extension to a finitely additive, nonnegative real valued set function on the class of all subsets of Ω. This is in contrast to the fact that finite additivity is superior to countable additivity.

8. If the sequence of bounded linear operators A_n on a Banach space converges pointwise to the finite-valued linear operator A, show that A is bounded; in fact
$$\|A\| \le \liminf_{n\to\infty} \|A_n\| < \sup_n \|A_n\| < \infty.$$

9. Let $\{A_\alpha : \alpha \in \Omega\}$ be a family of continuous linear operators from the Banach space X to the normed linear space M. Assume that A_α are weakly bounded, that is, $\sup_\alpha |f(A_\alpha x)| < \infty$ for all $x \in X$ and all $f \in M'$. Show that A_α are uniformly bounded, that is, $\sup_\alpha \|A_\alpha\| < \infty$.

10. (a) If the elements $x, y \in X$ belong to the normed linear space X, and $\sup_\alpha |f(x)| < \infty$ for every $f \in X'$, show that $\sup_\alpha \|x\| < \infty$.

    (b) If A is a linear operator from the normed linear space X to the normed linear space M, and $f \circ A$ is continuous for every $f \in M'$, show that A is continuous.

11. Let X, M, and N be normed linear spaces with X and M complete, and let $B : X \times M \to N$ be a bilinear form that is $B(x, y)$ is linear in x for each fixed y and linear in y for each fixed x. If for each $f \in N'$, $f(B(x, y))$ is continuous in x for each fixed y, and continuous in y for each fixed x, show that B is bounded. That is,
$$\sup\{|B(x, y)| : \|x\| \le 1, \|y\| \le 1\} < \infty.$$
(Apply Problem 9, for some positive number δ, to the family $\{\delta^{-2} B(\delta x, \delta y) : \|x\| \le 1, \|y\| \le 1\}$.)

12. Give an example of a linear unbounded operator from one normed linear space to another.

13. Let A be a bounded linear operator from the Banach space X onto the Banach space M. Show that there is a positive number δ such that for each $y \in M$ there is an $x \in X$ with $y = Ax$ and $\|x\| \le \delta \|y\|$. This theorem is sometimes called the solvability theorem.

3.5 Exercises

There is a vast literature on functional analysis, and we give only a few representative titles. Readable introductory treatments are given in Liusternik and Sobolev (1961), Taylor (1958), Rudin (1966), and Halmos (1951); the last deals extensively with Hilbert spaces. Among the more advanced treatments, Dunford and Schwartz (1958, 1963, 1971) emphasize normed spaces, Kelley and Namioka (1963) and Schaefer (1966) give precise topological vector space results, Yosida (1968) gives a broad survey of applications to differential equations, semigroup theory, and ergodic theorems, and Zeidler (1986–1990) gives a comprehensive treatment of nonlinear functional analysis.

Many exercises are by Conway (1978) and Wojtaszczyk (1991).

CHAPTER

4

BASIC CONCEPTS OF PROBABILITY

4.1 INTRODUCTION

The starting point for probability theory is a set Ω called the sample space whose points are in one-to-one correspondence with the possible outcomes of a given performance of a random experiment. For example, if we toss the head, we may take Ω to have two points, one for each natural possible, ω_i ($i = 1, \ldots, 6$). The sample space for a given experiment is not unique. For example, if two dice are tossed and N is the sum of the faces, we may take Ω to consist of 11 points corresponding to the outcomes $\omega = 2, 3, \ldots, 12$. The particular sample space to be used will be determined by the problem at hand. For example, in the dice-tossing example above, if we are interested in the sum of the two faces, the sample space corresponding to $N = 2, 3, \ldots, 12$ will not be adequate.

An "event" in a random experiment corresponds to a question whose answer is one of "yes" or "no" for example, in the tossing example above, ω_3 is the second natural number $\{\omega_i, i, j = 1, \ldots, 6\}$. We may ask the question "Is the maximum of the two numbers 3 and j for $j = 1, \ldots, 6$?" The subset of Ω associated with a "yes" answer is $A = \{(3, 1), (3, 2), (3, 3), (1, 3), (2, 3)\}$; the subset associated with a "no" answer is A^c.

Thus it is natural to define an event as a subset of Ω. However, in some situations not all subsets may be considered as events. As an example, admittedly somewhat artificial, suppose that a coin is tossed four times and Ω consists of the 16 sequences of length 4 with components H and T. Assume that only the results of the first two tosses are written down. If A is the set of points of Ω corresponding to at least three heads, then A is not "measurable," that is, the given information concerning ω is not sufficient to determine whether $\omega \in A$. More serious problems arise when Ω is not finite. In this case we can almost always find, by mathematical consistency requirements, it is also the event class to be the "Borel sets" that is the collection of all subsets of Ω.

The development of the mathematical theory will be facilitated if we require that the event class form a σ-field. This we may now formulate in terms

countable intersections, and complements of events and be assured that the resulting sets are also events.

Finally, we must talk about the probability $P(A)$ assigned to an event A. The basic empirical requirement is that $P(A)$ be approximately the relative frequency of A in a very large number of independent repetitions of the random experiment. It follows that P should be a nonnegative, countably additive set function, with $P(\Omega) = 1$. In order to be able to calculate the probability of a limit of events, we must require P to be continuously additive.

The above discussion may be summarized by saying that P is a probability measure on the σ-field $\mathcal{F}$. Thus the basic mathematical object we are to study is a probability space $(\Omega, \mathcal{F}, P)$.

4.2 Elementary Theorems of Set Theory

If the sample space Ω is a finite or countably infinite set, measure-theoretic difficulties do not arise. We take $\mathcal{F}$ to consist of all subsets of Ω, and assign probabilities in the following manner. Let $\Omega = \{\omega_1, \omega_2, \ldots\}$, and let $p_1, p_2, \ldots$ be nonnegative numbers whose sum is 1. If A is any subset of Ω, we define

$$P(A) = \sum_{\omega_i \in A} p_i.$$

In particular,

$$P(\{\omega_i\}) = p_i.$$

Then P is a probability measure, and the probability of the event A is computed simply by adding the probabilities of the individual points in A.

If Ω is countable, $\mathcal{F}$ consists of all subsets of Ω, and P is defined as above, then $(\Omega, \mathcal{F}, P)$ is said to be a discrete probability space.

Classical probability theory was concerned with the special case $\Omega = \{\omega_1, \ldots, \omega_n\}$, $p_1 = p_2 = \cdots = p_n = 1/n$. In this case,

$$P(A) = \frac{\text{number of points in } A}{\text{number of points in } \Omega}.$$

Thus to find $P(A)$ we count the number of favorable outcomes and divide by the total number of outcomes.

Unless otherwise specified, if Ω is countable, we always take $\mathcal{F}$ to consist of all subsets of Ω. Thus all subsets of Ω are events, all functions on Ω are measurable, and measure-theoretic machinations are not needed.

4.3 Basic Properties

Intuitively, two events A and B are independent if a statement concerning the occurrence or nonoccurrence of one of the events does not change the

odds about the other event. Let us translate the physical requirement into mathematical terms. Suppose, for example, that $P(A) = 0.6$ and $P(B) = 0.2$. In a long sequence of repetitions of the random situation, A will occur approximately 60% of the time. If B is to be independent of A, its occurrence or A will not influence the information B as the if we encounter only those trials on which A occurs, B will occur roughly 20% of the time; hence $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$. Similarly, the nonoccurrence of A will not influence the odds about B; hence $P(A^c \cap B) = P(A^c|B)P(B) = P(A^c)P(B)$.

Conversely, if $P(A \cap B) = P(A)P(B)$ and $P(A^c \cap B) = P(A^c)P(B)$, then B will be independent of A. If we examine only the trials on which A occurs, B will occur roughly 100% of the time in order to have $P(A \cap B) = P(A)P(B)$. Thus the assumption of A does not influence the odds about B; similarly, the nonoccurrence of A does not change the odds about B.

The above discussion suggests that we call the event B independent of A if and only if $P(A \cap B) = P(A)P(B)$ and $P(A^c \cap B) = P(A^c)P(B)$. However, the first condition implies the second. Suppose $P(A \cap B) = P(A)P(B)$. Then

$$\begin{aligned} P(A^c \cap B) &= P(B \quad A) = P(B - A \cap B) \\ &= P(B) - P(A \cap B) \qquad \text{since } A \cap B \subset B \\ &= P(B) - P(A)P(B) \qquad \text{by hypothesis} \\ &= (1 - P(A))P(B) \\ &= P(A^c)P(B). \end{aligned}$$

Therefore B is independent of A if and only if $P(A \cap B) = P(A)P(B)$. But this condition is symmetrical in A and B; it says that if A is independent of B, then B is independent of A if and only if A is independent of B. We may therefore summarize the definition of independence as follows.

4.8.1 *Definition* Two events A and B are independent iff $P(A \cap B) = P(A)P(B)$.

We now consider independence of more than two events. If the events $A_1, \ldots, A_n$ are to be independent, and if $i_1, \ldots, i_k$ are distinct indices, a statement about the occurrence or nonoccurrence of the events $A_{i_1}, \ldots, A_{i_k}$ should not change the probability of the remaining events. The proper generalization of the preceding definition is the following requirement.

4.8.2 *Definition* The events $A_1, \ldots, A_n$ in a given probability space are independent iff for all distinct collections $i_1, \ldots, i_k$ of distinct indices in $1, \ldots, n$ we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

4.3.3 *Comments.* (a) If the events A_i, $1 \le i \le n$, are independent and any event is obtained by taking unions, intersections or complements of the A_i, then a similar result holds. For example,

$$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n),$$

where $B_1, \ldots, B_n$ are distinct subsets and B_j is either A_j or A_j^c, $j = 1, \ldots, n$.

We have essentially proved this in the discussion preceding the definition of independence by showing that $P(A \mid B) = P(A)$ implies $P(A \cap M) = P(A)P(M)$.

(b) If $A_1, \ldots, A_n$ are events such that $A_1, \ldots, A_k$ are independent for all distinct indices $1, \ldots, k$, $k = 2, \ldots, n-1$, it does not follow that $A_1, \ldots, A_n$ are independent. For example, let a coin be tossed twice and assign probability $\frac{1}{4}$ to each of the outcomes HH, HT, TH, TT. Let $A =$ [1st toss is a head], $B =$ [2nd toss is a head], $C =$ [1st toss = 2nd toss] $=$ [HH, TT]. Then

$$P(A \cap B) = \frac{1}{4} = P(A)P(B),$$

$$P(A \cap C) = \frac{1}{4} = P(A)P(C),$$

$$P(B \cap C) = \frac{1}{4} = P(B)P(C),$$

but

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C).$$

Thus A and B are independent as are A and C, and also B and C, but A, B, and C are not independent.

Conversely, if $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$, it does not follow that $P(A_1 \mid A_2 \cap \cdots \cap A_n) = P(A_1 \mid A_2) \cdots P(A_1 \mid A_n)$ (why?). For example, suppose that two dice are tossed and let Ω be the 11 ordered pairs (i, j), $i, j = 1, 2, \ldots, 6$, with probability $\frac{1}{11}$ assigned to each point. Let

$$A = [\text{sum of dice} = 7 \text{ or } 11],$$

$$B = [\text{sum of dice} = 4, 5 \text{ or } 6],$$

$$C = [\text{the sum of the faces is } 6].$$

Then

$$P(A \cap B) = \frac{2}{11} = P(A)P(B) = \frac{2}{11},$$

$$P(A \cap C) = \frac{1}{11} = P(A)P(C) = \frac{1}{11},$$

$$P(B \cap C) = \frac{1}{11} = P(B)P(C) = \frac{1}{11},$$

but

$$P(A \cap B \cap C) = \frac{1}{11} = P(A)P(B)P(C).$$

4.4 Bernoulli trials

A sequence of *Bernoulli trials* consists of a finite or infinite number of independent observations with the property that each observation has only two possible results, called "success" and "failure". The probability of success on a given trial is p, and the probability of failure is $q = 1 - p$.

To construct a probability space that models the given physical situation, we take Ω to be the set of all finite sequences of length n with components S and F, with S representing success and F failure. Consider a typical sample point ω with S occurring in positions $i_1, \ldots, i_k$ and F occurring in the remaining positions $j_1, \ldots, j_{n-k}$. That is, the event H_i obtaining a success on trial i, or $H_i = \{\omega : \omega_i = S\}$, consists of all sample points with i as one of the coordinates $i_1, \ldots, i_k$. We have

$$\omega = A_{i_1} \cap \cdots \cap A_{i_k} \cap A'_{j_1} \cap \cdots \cap A'_{j_{n-k}}.$$

Since the trials are independent and $P(H_i) = p$ for all i, the probability assigned to ω is the product $p^k q^{n-k}$; it must be

$$P(\omega) = P(A_{i_1}) \cdots P(A_{i_k}) P(A'_{j_1}) \cdots P(A'_{j_{n-k}}) = p^k q^{n-k}.$$

Any two sequences $\omega, \omega' \in \Omega$ will assign the same probability to ω having exactly k terms, because such a sequence is determined by selecting k positions out of n for the successful terms. Thus the probability of obtaining exactly k successes is

$$p(k) = \binom{n}{k} p^k q^{n-k} \qquad k = 0, 1, \ldots, n. \tag{1}$$

By the binomial theorem, $\sum_{k=0}^{n} p(k) = (p + q)^n = 1$; therefore the sum of the probabilities assigned to all points is 1, and we have a legitimate probability assignment.

A sequence of *generalized Bernoulli trials* consists of n independent observations, each of which has r possible outcomes (say $A_1, \ldots, A_r$). If the k outcomes are labeled $1, \ldots, k$, the probability that A_i will occur on a given trial is p_i, where the p_i are nonnegative and $\sum_{i=1}^{r} p_i = 1$.

To construct an appropriate probability space, we take Ω to be the set of all ordered sequences of length n with components $a_1, \ldots, a_r$. If ω is a sample point having n_1 occurrences of A_1, n_2 of A_2, $\ldots$, n_r of A_r, the independence of the trials and the assumption that the probability of obtaining A_i on a given trial is p_i leads us to assign to ω the probability $p^{n_1} p^{n_2} \cdots p^{n_r}$.

Now to find the number of sequences in Ω in which A_1 occurs exactly n_1 times, A_2 occurs exactly n_2 times, $\ldots$, A_r occurs exactly n_r times, where $n_1 + \cdots + n_r = n$. Such a sequence is determined by selecting n_1 positions out of n to be occupied by A_1's, then n_2 positions out

from the remaining $n - n_1$ for the A_2's, and so on. Thus the total number of occurrences is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r}\binom{n_r}{n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}$$

The total probability obtained in (7) equals

$$\sum_{n_1!n_2!\cdots n_r!} \frac{n!}{n_1!n_2!\cdots n_r!} p_1^{n_1} p_2^{n_2}\cdots p_r^{n_r},$$

where the sum is taken over all nonnegative integers $n_1, \ldots, n_r$ whose sum is n. But this sum is $(p_1 + \cdots + p_r)^n = 1^n = 1$ by the multinomial theorem.

The probability that p_1 will occur exactly n_1 times, p_2 exactly n_2 times, and p_r will occur n_r times, is

$$p(n_1, n_2, \ldots, n_r) = \frac{n!}{n_1!n_2!\cdots n_r!} p_1^{n_1} p_2^{n_2}\cdots p_r^{n_r} \tag{7}$$

where $n_1, n_2, \ldots, n_r = 0, 1, 2, \ldots$ and $n_1 + n_2 + \cdots + n_r = n$.

4.5 Conditional Probability

In two events A and B are independent, a statement about the occurrence or nonoccurrence of one of the events does not change the odds about the other. In the absence of independence, the odds are altered, and the concept of conditional probability gives a quantitative measure of the change.

For example, suppose that the probability of A is $\frac{1}{6}$ and the probability of $A \cap B$ is $\frac{1}{12}$. If we repeat the experiment independently a large number of times and examine only the trials in which A has occurred, B will occur roughly $\frac{1}{12}$ of the time. In general, the ratio $P(A \cap B)/P(A)$ is a measure of the probability of B under the assumption that the event A is known to have occurred.

We therefore define the conditional probability of B given A as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{1}$$

provided $P(A) > 0$.

In the next chapter, we shall discuss in detail the concept of conditional probability $P(B \mid A)$ when the event A has probability 0. This is not a degenerate case that can safely be neglected in many examples. Of course, the denominator in (1) no longer makes sense, and the approach will be somewhat indirect. At this point we shall only derive a few consequences of (1).

4.5.1 *Theorem.* (a) If $P(B) > 0$, A and B are independent if $P(A \mid B) = P(A)$. Similarly, independence is equivalent to $P(A \mid B) = P(A) P(B) + P(A^c) P(B^c)$. (b) If $P(A_1 \cap \cdots \cap A_{n-1}) > 0$, then

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Proof. (a) This follows from the definitions of independence and conditional probability. (b) If $P(A_1 \cap \cdots \cap A_{n-1}) > 0$ then $P(A_1 \cap \cdots \cap A_{n-2}) > 0$, $P(A_1) > 0$, so all conditional probabilities are well defined. Now by the definition of conditional probability,

$$P(A_1 \cap \cdots \cap A_n) = P(A_1 \cap \cdots \cap A_{n-1}) P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

An induction argument completes the proof. □

The following result will be quite useful.

4.5.2 *Theorem of Total Probability.* Let $B_1, B_2, \ldots$ be a finite or countably infinite family of mutually exclusive and exhaustive events; that is, the B_i are disjoint and their union is Ω.

(a) Let A be any event. Then $P(A) = \sum_i P(A \mid B_i)$. Thus $P(A)$ is calculated by breaking A into mutually exclusive pieces in various ways in which A can happen, and adding the individual probabilities.

(b) $P(A) = \sum_i P(B_i) P(A \mid B_i)$, where the sum is taken over those i for which $P(B_i) > 0$. Thus $P(A)$ is a weighted average of the conditional probabilities $P(A \mid B_i)$.

Proof.

(a) $P(A) = P(A \cap \Omega) = P(A \cap (\bigcup_i B_i)) = P(\bigcup_i (A \cap B_i)) = \sum_i P(A \cap B_i)$.

(b) This follows from (a) and the fact that $P(A \cap B_i) = 0$ if $P(B_i) = 0$, and equals $P(B_i) P(A \mid B_i)$ if $P(B_i) > 0$. □

4.5.3 *Example.* A positive integer n is selected with $P(N = n) = (\frac{1}{2})^n$, $n = 1, 2, \ldots$. If $N = n$, a coin with probability p of heads is tossed n times. Find the probability that the resulting tosses are all heads.

Solution. We have specified $P(B_n) = (\frac{1}{2})^n$, where $B_n = [N = n]$, $n = 1, 2, \ldots$. If A is the event that the coin comes up heads each time, we have specified $P(A \mid B_n)$. By the theorem of total probability, it is enough to determine $P(A \cap B_n)$. However, we may take Ω to consist of all ordered pairs (n, k), $n = 1, 2, \ldots$, $k = 0, 1, \ldots, n$. Let P assign to the point (n, k) the probability $\binom{n}{k} p^k (1-p)^{n-k} (\frac{1}{2})^n$.

and to ω_2 the probability $(\frac{1}{2})^2(1-p)^{2-2}$. We may then verify that $P(\{\omega_1\}) = (\frac{1}{2})^2$, $P(\{\omega_2\}) = p^2$, $P(\{\omega_3\}) = p(1-p)$. Hence

$$P(\{\omega\}) = \sum_{\omega \in \{\omega_1,\omega_2,\omega_3\}} (\tfrac{1}{2})^{2-r} p^r (1-p)^{2-r} = \frac{(\frac{1}{2})^2 p^2}{1-(\frac{1}{2})^2 (1-p)^2}.$$

4.6 Random variables

In many applications in which probability theory is used, we are concerned with a numerical quantity that is associated in some way with a random experiment. If $(\Omega, \mathcal{F}, P)$ is a probability space, the outcome ω of the experiment corresponds to the point $\omega \in \Omega$, a numerical quantity associated with the experiment is then a number $X(\omega)$. Thus X is a function from the sample space Ω to the real numbers. For example, if $(\Omega, \mathcal{F}, P)$ corresponds to a sequence of four Bernoulli trials (4.5) and X is the number of successes, then $X(\omega_1) = 4$, $X(\omega_2) = 2$, $X(\omega_3) = 3$, and so on.

In most instances in which a random variable is defined on a given probability space, we generally want to know the probability of events involving X. For example, the probabilities that in a given person each ω in the experiment the value of X will belong to B, where B is a set of real numbers. In particular, we will be interested in $P[\{\omega : a \le X(\omega) \le b\}]$ for all real a, b. Thus $\mathcal{F}$ must contain all sets of the form $X^{-1}[a, b]$, and therefore all sets $X^{-1}(B)$, B a Borel set in $\mathbb{R}$.

4.6.1 *Definition.* A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a Borel measurable function from Ω to $\mathbb{R}$. In the terminology of 1.3, we have $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In many situations it is convenient to allow X to take on the values $\pm\infty$. X is said to be an extended random variable if X is a Borel measurable function from Ω to $\bar{\mathbb{R}}$ that is $X : (\Omega, \mathcal{F}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$.

If X is a random variable on $(\Omega, \mathcal{F}, P)$ the *probability measure induced by* X is the probability measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by

$$P_X(B) = P[\{\omega : X(\omega) \in B\}], \qquad B \in \mathcal{B}(\mathbb{R}).$$

The numbers $P_X(B)$, $B \in \mathcal{B}(\mathbb{R})$, completely characterize the random variable X in the sense that they specify the probabilities of all events involving X. It is useful to know that this information may be captured by a single function from $\mathbb{R}$ to $\mathbb{R}$.

4.6.2 *Definition.* The *distribution function* of a random variable X is the function $F = F_X$ from $\mathbb{R}$ to $[0, 1]$ given by

$$F(x) = P[\{\omega : X(\omega) \le x\}], \qquad x \text{ real}.$$

Since, for $a < b$, $F(a) = P[\omega : X(\omega) \le a] \le P[\omega : X(\omega) \le b] = F(b)$, F is a distribution function corresponding to the Lebesgue-Stieltjes measure P_X on $(R^1, \mathcal{B}^1)$. In particular, F is nondecreasing and right-continuous, $F(x) \to 1$ as $x \to \infty$, and $F(x) \to 0$ as $x \to -\infty$. There may be distribution functions corresponding to P_X; we choose the one with $F(x) = P_X(-\infty, x]$.

Many of the following statements remain valid if X is a random variable with distribution function F, where F is a given function from R^1 to $[0, 1]$ that is nondecreasing and right-continuous with $F(-\infty) = 0$, $F(\infty) = 1$. There is no reference to the underlying probability space $(\Omega, \mathcal{F}, P)$, and actually the nature of the underlying space is unimportant. The distribution function F determines the probability measure P_X, which in turn determines the probability of all events involving X. The only thing we have to check is that there does exist some $(\Omega, \mathcal{F}, P)$ on which a random variable X with a given distribution function F can be defined. In fact, we can always supply the probability space in a canonical way: take $\Omega = R^1$, $\mathcal{F} = \mathcal{B}^1$, $P = P_X$, with P_X the Lebesgue-Stieltjes measure corresponding to F, and define $X(\omega) = \omega$, so that P_X is the identity map. Since $P_X(B) = P[\omega : X(\omega) \in B] = P_X(B)$, X has indeed probability measure P_X and therefore distribution function F.

Thus if $F : R^1 \to [0, 1]$ is nondecreasing and right-continuous with $F(-\infty) = 0$ and $F(\infty) = 1$, then F is the distribution function of some random variable.

We isolate some particularly important classes of random variables.

4.6.3 Discrete and Continuous. Let X be a random variable on $(\Omega, \mathcal{F}, P)$. We say that X is *simple* if $X(\Omega)$ can assume only finitely many possible values; X is *discrete* if $X(\Omega)$ is countable. (Any random variable on a discrete probability space is discrete, since Ω is countable.)

If X is discrete and the values $\{x_n\}$ of X can be arranged so that $x_1, x_2, \ldots$ are all of the points of positive increase of F, then F is a step function with a jump of magnitude $p_n = P[X = x_n]$ at each x_n of magnitude $p_n = P[X = x_n]$. It is natural to associate the x_n that are the jump points of F with the values that X actually assumes. To see this, observe that if $x_{n-1} < x < x_n$, then $F(x) - F(x_{n-1}) = P[\omega : X(\omega) \le x] - P[\omega : X(\omega) \le x_{n-1}] = P[\omega : x_{n-1} < X(\omega) \le x] = 0$; if x_n is the nth jump point, then $F(x_n) - F(x_{n-1}) = p_n$.

If X is an arbitrary discrete random variable, the properties of X are completely determined by the probability function p_X defined by

$$p_X(x) = P[X = x], \qquad -\infty < x < \infty.$$

Explicitly,

$$P_X(B) = \sum_{x \in B} p_X(x).$$

This is a countable sum since p_X is 0 except at the x_n.

Thus a discrete random variable can be specified by a countable set $\{x_1, x_2, \ldots\}$ of real numbers and a set of probabilities $\{p_1, p_2, \ldots\}$ with $p_i = P(X = x_i) \geq 0$ and $\sum_i p_i = 1$, where p_i is interpreted as $P[X = x_i]$ = the probability that X belongs to x_i is found by assigning the p_i for those x_i which belong to B.

The random variable X is said to be absolutely continuous if there is a nonnegative real-valued Borel measurable function f on $\mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \qquad x \in \mathbb{R}.$$

Then f is called a density function for X. Since $F(+\infty) = 1$, we must have $\int_{-\infty}^{\infty} f(t)\,dt = 1$.

If X is absolutely continuous with density f, it follows that

$$P_X(B) = \int_B f(t)\,dt \qquad \text{for each} \qquad B \in \mathcal{A}(\mathbb{R}).$$

For the measure ν defined by $\nu(B) = \int_B f(t)\,dt$, $B \in \mathcal{B}(\mathbb{R})$, satisfies $\nu(a, b] = F(b) - F(a)$, $a < b$. Thus ν and P_X coincide on intervals corresponding to $\mathcal{I}$, hence $\nu = P_X$.

The absolute continuity of X means that $P_X \ll$ Lebesgue measure, so by 3.5.1, F_X is an absolutely continuous function.

Any nonnegative Borel measurable function f on $\mathbb{R}$ with $\int_{-\infty}^{\infty} f(t)\,dt = 1$ is the density of some absolutely continuous random variable X. For $g(x) = \int_{-\infty}^{x} f(t)\,dt$, g is clearly nondecreasing, and by 3.5.1, g is absolutely continuous, hence continuous, on $\mathbb{R}$. Since $g(-\infty) = 0$, $g(+\infty) = 1$, g is the distribution function of some random variable X, and f must be its density.

The random variable X is said to be continuous if its distribution function F is continuous on all of $\mathbb{R}$. Equivalently, X is continuous if $P[X = x] = 0$ for all x, hence $P[B] = 0$ for all countable $B \in \mathcal{A}(\mathbb{R})$.

We have seen that absolute continuity implies continuity. If f is the Cantor function (Problem 3, section 2.3), extended so that $f(x) = 0$ for $x \leq 0$ and $f(x) = 1$ for $x \geq 1$, then if X is a random variable with distribution function $F = f$, X is continuous but not absolutely continuous.

Some typical examples of density functions are:

(1) Uniform density on $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

(2) Exponential density:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

where $\lambda > 0$.

(j) Two sided exponential density:

$$f(x) = \frac{1}{2}\alpha e^{-\alpha|x|}, \qquad \alpha > 0, \; -\infty < x < \infty.$$

(k) Normal density:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \quad \sigma > 0, \; -\infty < m < \infty, \; -\infty < x < \infty.$$

(l) Cauchy density:

$$f(x) = \frac{\theta}{\pi(x^2+\theta^2)}, \qquad \theta > 0, \; -\infty < x < \infty.$$

Finally, some remarks on terminology. We will often abbreviate $\mu(A) = 0$ by $\mu[X \in A] = 0$; note that this set is also $X^{-1}(A)$, the pre-image of A under the mapping X. Similarly, $\mu[X \in A] = 1$ will be abbreviated by $\mu[X \in A^c] = 0$. The letter $\mathcal{X}$ will always stand for a finite-dimensional Euclidean space. Thus $\mu \circ X^{-1}: \mathcal{B}^n \to [0,1]$ means that $\mu^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$ for each $B \in \mathcal{B}(\mathbb{R}^n)$. The phrase "almost surely," abbreviated a.s., is often used in the literature. It means "almost everywhere" with respect to a specified probability measure.

4.7 Random Vectors

We now consider situations involving more than one random variable associated with the same experiment.

An n-dimensional random vector on a probability space $(\Omega, \mathcal{F}, P)$ is a Borel measurable mapping from Ω to $\mathbb{R}^n$.

If $X: \Omega \to \mathbb{R}^n$ and $X_j = p_j \circ X$, where p_j is the projection of $\mathbb{R}^n$ onto the j-th coordinate space, then X is Borel measurable if each X_j is Borel measurable (see 1.5.8). Thus a random vector may be regarded as an n-tuple $(X_1, \ldots, X_n)$ of random variables.

Much of the development of the previous section carries over. As before, the probability measure induced by the random vector X is defined by

$$P_X(B) = P[\omega: X(\omega) \in B], \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

The cumulative distribution function of X is the function $F: \mathbb{R}^n \to [0,1]$ defined by

$$F(x) = P_X[(-\infty, x]] = P[\omega: X_j(\omega) \le x_j, \quad j = 1, \ldots, n]$$

F is also called the joint distribution function of $X_1, \ldots, X_n$. F is increasing and right-continuous on $\mathbb{R}^n$, and P_X is the Lebesgue–Stieltjes measure determined by F (see 1.4).

By 1.9.7 we have

$$P[a_1 < X_1 \le b_1, \ldots, a_n < X_n \le b_n] = \begin{cases} 1 & \text{if } a_i < b_i \text{ for all } i \\ 0 & \text{if } a_i \ge b_i \text{ for some } i \end{cases} \qquad (1)$$

for any n-dimensional rectangle $[a, b]$.

If F is a distribution function on $\mathbb{R}^n$ that satisfies (1), then F is the distribution function of some random n-vector X, and the underlying probability space can be constructed in a canonical way. Take $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}^n$, $\mathcal{B}^n$ the σ-algebra of Borel subsets of $\mathbb{R}^n$, and X the identity mapping on Ω. Since $P_X(B) = P[X \in B]$, X has induced probability measure P_X. Furthermore the distribution function of X is

$$\begin{aligned} F_X(x) = P_X[(-\infty, x]] &= P[(-\infty, x]] \\ &= \lim_{y \to x} P[(-\infty, y]] && \text{by 1.4.5} \\ &= \lim_{y \to x} F(y) && \text{since } P \text{ is the Lebesgue–Stieltjes measure determined by } F \\ &= F(x) && \text{by 1.4.5(ii) and the hypothesis that } F(y) \to 0 \text{ if any } y_i \to -\infty \end{aligned}$$

[The hypothesis that $F(x) \to 1$ as $x_i \to \infty$ implies that P is actually a probability measure.]

The random vector X is said to be discrete if the set of values of X is finite or countably infinite, or equivalently, if the components $X_1, \ldots, X_n$ are all discrete. In this case the properties of X are determined by the probability function p, given by $p(x) = p[X = x] = P[X_1 = x_1, \ldots, X_n = x_n]$ and $P[X \in B] = \sum_{x \in B} p(x)$.

The random vector X is said to be absolutely continuous if there is a nonnegative Borel measurable function f on $\mathbb{R}^n$, called the density function of X, such that

$$F(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t)\, dt, \qquad x = (x_1, \ldots, x_n).$$

It follows that

$$P[X \in B] = \int_B f(x)\, dx \qquad \text{for all} \qquad B = (a_1, b_1] \times \cdots \times (a_n, b_n]$$

or the measure μ created by $\mu(B) = \int_B f(x)\, dx$, $B \in \mathcal{B}(\mathbb{R}^n)$, satisfies $\mu((a, b]) = \int_{(a,b]} f(x)\, dx = F(a, b]$ [see 1.4.1(iii)]. Thus μ is the Lebesgue–Stieltjes measure determined by F, hence $\mu = P_X$.

Just as in the previous section, any nonnegative Borel measurable function f on $\mathbb{R}^n$ with $\int_{\mathbb{R}^n} f(x)\,dx = 1$ is the density of some absolutely continuous random vector X.

4.5 Independent Random Variables

We now wish to look precisely at the notion of independence of events as it applies to independent random variables. Intuitively, independence of $X_1, \ldots, X_n$ means that no statement about one or more of the X_i does not affect the probabilities concerning the remaining X_i. Now a statement about X_i corresponds to an event of the form $A_i = \{\omega : X_i(\omega) \in B_i\}$; thus the events $A_1, \ldots, A_n$ will be independent. The formal definition is as follows.

4.5.1 Definition. Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, P)$. Then $X_1, \ldots, X_n$ are said to be independent iff for all sets $B_1, \ldots, B_n \in \mathcal{B}^1(\mathbb{R})$, we have

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1] \cdots P[X_n \in B_n].$$

By 2.6.8(b), independence of $X_1, \ldots, X_n$ may be expressed by saying that if $X = (X_1, \ldots, X_n)$, then P_X is the product of the P_{X_i}, $i = 1, \ldots, n$.

4.5.2 Comments. (a) If $X_1, \ldots, X_n$ are independent, so are $Y_1, \ldots, Y_n$ for $Y_i = u_i \circ X_i$, $i = 1, \ldots, n$, where $u_1, \ldots, u_n : (\mathbb{R}, \mathcal{B}^1(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}^1(\mathbb{R}))$. That is,

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1, \ldots, X_{n-1} \in B_{n-1}]\, P[X_n \in B_n]$$

$$= P[X_1 \in B_1] \cdots P[X_n \in B_n].$$

Thus it is not necessary to check all subcollections of the collection of events $\{X_i \in B_i\}$, $i = 1, \ldots, n$, as in the definition of independent events in 4.3.5.

(b) The independence of arbitrary random vectors is defined exactly as above, with $\mathcal{B}^n(\mathbb{R})$ replacing $\mathcal{B}^1(\mathbb{R})$. In fact, a random object, that is a map $X : \Omega \to \Omega_1$, $\mathcal{F} \to \mathcal{G}_1$, $X(\omega)$ where Ω is an arbitrary set and $\mathcal{F}$ is a field of subsets of Ω. Then the $X_1, \ldots, X_n$ are said to be independent iff for all $B_1 \in \mathcal{F}_1, \ldots, B_n \in \mathcal{F}_n$ we have

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1] \cdots P[X_n \in B_n].$$

(c) Let $\{X_i : i \in I\}$ be an arbitrary indexed set of random objects. The X_i are said to be independent iff $X_{i_1}, \ldots, X_{i_n}$ are independent for each finite set $\{i_1, \ldots, i_n}$ of distinct indices in I.

(d) If the $X_i : (\Omega, \mathcal{F}) \to (\Omega_i, \mathcal{F}_i)$ are independent random objects and $g_i : (\Omega_i, \mathcal{F}_i) \to (\Omega_i', \mathcal{F}_i')$, then the random objects $g_i \circ X_i : \Omega \to \Omega_i'$ are independent.

("Functions of independent random objects are independent.") This follows from the definition of independence and the fact that

$$[g_i(X_i) \in H_i] = [X_i \in g_i^{-1}(H_i)].$$

Independence of random variables may be characterized in terms of distribution functions as follows.

4.8.1 *Theorem.* Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, P)$, let F be their distribution function on $\mathbb{R}^n$, and let F_i be the distribution function of X_i ($i = 1, \ldots, n$). Then $X_1, \ldots, X_n$ are independent iff

$$F(x_1, \ldots, x_n) = F_1(x_1) F_2(x_2) \cdots F_n(x_n) \quad \text{for all real } x_1, \ldots, x_n.$$

Proof. If $X_1, \ldots, X_n$ are independent, then

$$F(x_1, \ldots, x_n) = P[X_1 \le x_1, \ldots, X_n \le x_n] = \prod_{i=1}^{n} P[X_i \le x_i] = \prod_{i=1}^{n} F_i(x_i).$$

Conversely, assume $F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)$ for all $x_1, \ldots, x_n$. Then

$$P[a_1 < X_1 \le b_1, \ldots, a_n < X_n \le b_n] = \sum_{a_1}^{b_1} \cdots \sum_{a_n}^{b_n} [F(x_1, \ldots, x_n) - F(x_1, \ldots, x_{n-1}, a_n)] - \prod_{i=1}^{n} [F_i(b_i) - F_i(a_i)]$$

(see (2.5.10)). Thus

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1] \cdots P[X_n \in B_n] \tag{*}$$

when the B_i are finite semiclosed intervals of reals. Now fix the intervals $B_2, \ldots, B_n$. The collection $\mathcal{M}$ of sets $M \in \mathcal{B}(\mathbb{R})$ for which $(*)$ holds is monotone class and, by the preceding remarks, contains all finite disjoint unions of right-semiclosed intervals and therefore $\mathcal{M} = \mathcal{B}(\mathbb{R})$ by the monotone class theorem. Applying the same reasoning to each coordinate in turn we obtain the independence of $X_1, \ldots, X_n$. (Similarly, we proved by induction that if $\Omega_1, \ldots, \Omega_n$ are arbitrary Borel sets and $\mathcal{G}_i$ is π-system of subsets of Ω_i containing Ω_i, then $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are independent if $\mathcal{G}_i \subset \mathcal{F}_i$ for each i.)

We may also characterize independence in terms of densities.

4.8.2 *Theorem.* Let $X = (X_1, \ldots, X_n)$ have a density f that is continuous on $\mathbb{R}^n$. Let f_i be the density of X_i. Then $X_1, \ldots, X_n$ are independent iff $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$

$= F_1(x_1) \cdots F_n(x_n)$ for all $(x_1, \ldots, x_n)$ except possibly on a Borel set of $\mathbb{R}^n$ with Lebesgue measure zero.

Proof.

$$F_1(x_1) = P(X_1 \le x_1) = P(X_1 \le x_1, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R})$$
$$= \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \ldots, t_n)\, dt_2 \cdots dt_n\, dt_1.$$

By definition of absolute continuity, X_1 has a density given by

$$f_1(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_2 \cdots dx_n$$

(integrate over x_1 = integrate over all variables). Borel measurability of f_1 follows from Fubini's Theorem. Similarly, each X_i has a density f_i obtained by integrating out all variables except x_i.

Now $f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$ a.e. Thus

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) dt_1 \cdots dt_n = F_1(x_1) \cdots F_n(x_n)$$

so by 4.5.3 the X_i are independent. Conversely, if the X_i are independent then

$$F_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n)$$
$$= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_1(t_1) \cdots f_n(t_n)\, dt_1 \cdots dt_n \qquad (1)$$

Thus (as the end of 4.5.3) if $g(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$ then

$$P_{\mathbf{X}}(B) = \int_B g(x)\, dx_1 \cdots dx_n \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

But $F_i(x) = \int_{-\infty}^{x} f_i(t)\, dt$, and it follows that $f_i = g_i$ a.e. (Take $x \to \infty$ in (1).) ∎

4.6.7 Corollary. If $X_1, \ldots, X_n$ are independent and X_i has density f_i ($i = 1, \ldots, n$), then $\mathbf{X}$ has density f given by $f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$.

Proof. Combine (1) of 4.6.6 with (1). ∎

If $X_1, \ldots, X_n$ each have a density, it does not follow that $(X_1, \ldots, X_n)$ has a density; this will be false without the independence hypothesis (see Problem 1).

Problems

1. Give an example of random variables X and Y (on the same probability space) such that X and Y each have densities, but (X, Y) does not.
2. Give an example to show that even if (X, Y) has a density, it is not determined by the individual densities of X and Y.
3. Let $X_1, \ldots, X_n$ be discrete random variables. Show that the X_i are independent if

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \quad \text{for all } x_1, \ldots, x_n$$

4. Let $(\Omega, \mathcal{F}, P)$ be a probability space. The classes $\mathcal{H}_1, \ldots, \mathcal{H}_n$ of events in $\mathcal{F}$ are said to be independent if, given any choice of $G_1 \in \mathcal{H}_1, \ldots, G_n \in \mathcal{H}_n$, the events $G_1, \ldots, G_n$ are independent. (Thus the random objects $X_i = (\Omega, \mathcal{H}_i, P)$ = "$(\Omega, \mathcal{H}_i, P)$" are independent if all the classes $\mathcal{H}_1, \ldots, \mathcal{H}_n$ are independent.)

   If the $\mathcal{H}_i$ ($\subset \mathcal{F}$) are independent, show that if the following sets are adjoined to each $\mathcal{H}_i$, the enlarged classes still remain independent.

   (a) Proper differences $A - B$, $A, B \in \mathcal{H}_i$, $B \subset A$
   (b) The space Ω
   (c) Countable disjoint unions of sets in $\mathcal{H}_i$
   (d) Limits of monotone sequences in $\mathcal{H}_i$

   Give an example to show that mere intersections cannot be treated.

   If you are familiar with Dynkin's lemma, show that if the $\mathcal{H}_i$ are independent classes each closed under finite intersection, the minimal σ-fields containing the $\mathcal{H}_i$ are also independent.

4.9 Some Measures More about Independence

In this section we give some illustrations to show how some of the concepts above can be elementarily probabilistically connected with the present measure theoretic framework.

4.9.1 *Example.* Three numbers X and Y are picked at random between 0 and 1. Assume that X and Y are independent and that each is uniformly distributed (that is, X and Y have densities f and g given by $f(x) = f(x-1) = 1$, $f(x) = 0$ otherwise). Let Z be the point in XY and let us find the distribution function of Z.

We take $\Omega = \mathbb{R}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$, $X(\omega, y) = \omega$, $Y(\omega, y) = y$. By 4.5.2, (X, Y) must have density $f(x, y) = f_1(x) f_2(y)$. Hence

$$P[(X, Y) \in B] = \iint_B f_1(x) f_2(y)\, dx\, dy.$$

Thus we have our probability measure P be

$$P(B) = \iint_B f_1(x) f_2(y)\, dx\, dy \qquad B \in \mathcal{B}(\mathbb{R}^2).$$

Now

$$F_Z(z) = P[Z \le z] = P[(X, Y) : X \le z] = \iint_{x \le z} f_1(x) f_2(y)\, dx\, dy.$$

Since X and Y are between 0 and 1 with probability 1, $f_1(x) = 1$ for $0 \le x \le 1$ and $f_1(x) = 0$ for $x \notin [0, 1]$. Since $f_2(y) = 1$ for $0 \le y \le 1$, $0 \le z \le 1$, and G is as pictured,

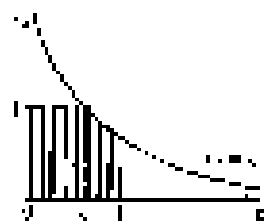

Figure 4.5.1

$F_Z(z)$ is the area of the shaded region in Fig. 4.5.1; that is, $F_Z(z) = \int_0^z (1 - \ln x)\, dx$, $0 \le z \le 1$. Thus Z has a density

$$f_Z(z) = F_Z'(z) = \begin{cases} -\ln z & 0 < z < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Note that f_Z is unbounded, as foretold, namely f_Z is always finite.

4.5.2 *Example.* Let X, Y, and Z be independent random variables, each normally distributed with $\mu = 0$, $\sigma = 1$, that is, X, Y, and Z each have the normal density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(x - m)^2}{2\sigma^2}\right] \qquad m \in \mathbb{R},\ \sigma > 0.$$

Let $W = (X^2 + Y^2 + Z^2)^{1/2}$ (take the positive square root so that $W \geq 0$). Find the distribution function of W.

Solution Take $\Omega = \mathbb{R}^3$, $\mathcal{F}$ = Borel sets, $X(\omega_1, \omega_2, \omega_3) = \omega_1$, $Y(\omega_1, \omega_2, \omega_3) = \omega_2$, $Z(\omega_1, \omega_2, \omega_3) = \omega_3$, and

$$P(B) = \iiint_B f(x, y, z)\, dx\, dy\, dz \qquad B \in \mathcal{B}(\mathbb{R}^3),$$

where

$$f(x, y, z) = f_X(x) f_Y(y) f_Z(z) = \frac{1}{(2\pi)^{3/2}} \exp\left[-\frac{1}{2}(x^2 + y^2 + z^2)\right].$$

Now

$$F_W(w) = P\{W \leq w\} = P\{X^2 + Y^2 + Z^2 \leq w^2\}.$$

If $w \geq 0$,

$$F_W(w) = \iiint_{x^2+y^2+z^2 \leq w^2} (2\pi)^{-3/2} \exp\left[-\frac{1}{2}(x^2 + y^2 + z^2)\right] dx\, dy\, dz$$

or in spherical coordinates,

$$F_W(w) = \int_0^{2\pi} d\theta \int_0^{\pi} d\phi \int_0^{w} (2\pi)^{-3/2} e^{-\rho^2/2} \rho^2 \sin\phi\, d\rho$$

$$= (2\pi)^{-1/2} \sqrt{\frac{2}{\pi}} \int_0^{w} \rho^2 \exp\left[-\frac{1}{2}\rho^2\right] d\rho.$$

Since W is obviously nonnegative, we have

$$f_W(w) = \begin{cases} \sqrt{\frac{2}{\pi}}\, w^2 \exp\left[-\frac{1}{2}w^2\right], & w \geq 0, \\ 0, & w < 0. \end{cases}$$

4.9.3 *Example.* Let $X_1, \ldots, X_n$ be independent random variables, each with density f and cumulative distribution function F; that is, $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, $X_k(\omega_1, \ldots, \omega_n) = \omega_k$, $1 \leq k \leq n$,

$$P(B) = \int_B f(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n \qquad B \in \mathcal{B}(\mathbb{R}^n).$$

Let R_k be the kth range variable, as in Section 4.7; for example, if $n = 4$, $X(\omega_1) = 2$, $X(\omega_2) = 1.4$, $X(\omega_3) = -7$, $X(\omega_4) = 8$, then $R_1(\omega_1) = \max X(\omega_i) - \min X(\omega_i)$

$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n = 0$ for $i \neq j$.

and therefore

$$P[X_i = X_j \text{ for at least one } i \neq j] \leq \sum_{i<j} P[X_i - X_j = 0] = 0.$$

This is known with probability 1 and can be ignored.

Now the joint cumulative distribution function of $(Y_1, \ldots, Y_n)$ and the joint distribution function of $(X_1, \ldots, X_n)$

Now

$$P[Y_j = y_j] = \sum_{i=1}^{n} P[X_i \leq x_i, Y_j = y_j] \qquad (y_j < y_{j+1}) \qquad (1)$$

and, for example,

$$P[Y_j \leq x, Y_j = y_j] = P[X \leq x \text{ exactly } j-1 \text{ of the random variables } X_1, \ldots, X_n \text{ are less than } y_j \text{ and the remaining } n-j \text{ random variables are greater than } Y_j]. \qquad (2)$$

But using Fubini's theorem

$$P[X \leq x, X_1 < x_1, \ldots, X_j < x_j, X_{j+1} = y_j, \ldots, X_n = y_j] = \int_{-\infty}^{x} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_j} \int_{y_j}^{\infty} \cdots \int_{y_j}^{\infty} f(x_1, \ldots, x_n)\, dx_n \cdots dx_1 = \int_{-\infty}^{x} f(x, \ldots, x, y_j, \ldots, y_j)\, dx \qquad (3)$$

Now by symmetry, (3) is the sum of $\binom{n}{j-1}$ terms, each one which has the same value as (3), since we may select the $j-1$ random variables to be less

can A_1 in $\binom{n}{1}$ ways. Also, each term in the summation (1) has the same value as (2). Thus

$$F_{Y_n}(y) = n\int_{-\infty}^{y} \binom{n-1}{0} [F_X(x)]^0 [1 - F_X(x)]^{n-1} f_X(x)\,dx$$

so that Y_n is absolutely continuous, with density

$$f_{Y_n}(y) = n\binom{n-1}{0} f_X(y)[1 - F_X(y)]^{n-1}.$$

We now find the joint distribution function of $Y_1, \ldots, Y_n$. Let $b_1 < b_2 < \cdots < b_n$. Then

$$\begin{aligned}
&P\{Y_1 \le b_1, \ldots, Y_n \le b_n\} \\
&\quad = n!P\{X_1 \le b_1, \ldots, Y_n \le b_n, X_1 < X_2 < \cdots < X_n\} \quad \text{by symmetry} \\
&\quad = n!P\{X_1 \le b_1, X_1 \le X_2 \le b_2, \ldots, X_{n-1} \le X_n \le b_n, X_1 < X_2 < \cdots < X_n\} \\
&\quad = n!\int_{-\infty}^{b_1} f(x_1)\,dx_1 \int_{x_1}^{b_2} f(x_2)\,dx_2 \cdots \int_{x_{n-1}}^{b_n} f(x_n)\,dx_n \\
&\quad = \int_{-\infty}^{b_1} \cdots \int_{-\infty}^{b_n} g(x_1, \ldots, x_n)\,dx_1 \cdots dx_n
\end{aligned}$$

where

$$g(x_1, \ldots, x_n) = \begin{cases} n!f(x_1) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases}$$

Thus $(Y_1, \ldots, Y_n)$ is absolutely continuous with density g. (Note that f_{Y_n} can be found from g (see 4.5.4), but the calculation is not any simpler than the direct method we have used above.)

4.6.4 Example. Let X be an absolutely continuous random variable with density f, assumed to be piecewise continuous. Let D be a Borel subset of $\mathbb{R}$ such that D includes the range of X, and let u be a Borel measurable function from D to $\mathbb{R}$.

If $Y = u(X)$, we have no way to graph the cumulative distribution function of Y. Discouraged as a general result, we say that we know the distribution of Y means that we know how to calculate $P\{Y \in B\}$ for all Borel subsets B of $\mathbb{R}$. Thus the distribution may be specified by giving the induced probability measure P_Y on the real line. Thus if f is absolutely continuous, as usually is appropriate, and u is discrete, the probabilistic function is clear. If $\Omega = \{\omega_1, \omega_2, \ldots\}$, $P\{\omega_i\} = p_i$, and X is an arbitrary random variable, the distribution of X means the probability measure P_X, defined by $P_X(B) = P\{X^{-1}(B)\} = P\{\omega_i : X(\omega_i) \in B\}$.

Assume that D is an open interval (a, b) and u is either strictly increasing or strictly decreasing, with inverse h. Assume also that u has a continuous

monotone derivative (hence we denote it by g). We show that Y is absolutely continuous with density

$$f_Y(y) = \begin{cases} f_X(h(y))\,|h'(y)|, & y \in g(S_X), \\ 0 & \text{otherwise.} \end{cases}$$

We can write for $y \in g(S_X)$,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) \\ &= P(X \le h(y)) \quad \text{if } g \text{ is increasing,} \\ &= P(X \ge h(y)) \quad \text{if } g \text{ is decreasing.} \end{aligned}$$

(see Fig. 4.9).

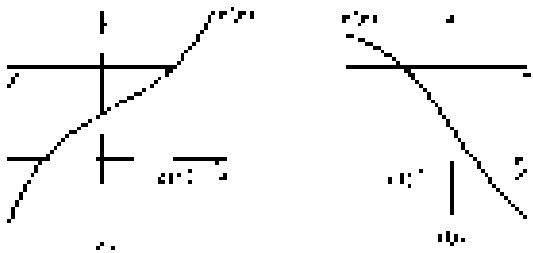

Figure 4.9: Illustrating the relation between the cdf's of X and $Y = g(X)$.

Thus

$$F_Y(y) = \begin{cases} \displaystyle\int_{-\infty}^{h(y)} f_X(x)\,dx & \text{if } g \text{ is increasing,} \\ \displaystyle\int_{h(y)}^{\infty} f_X(x)\,dx & \text{if } g \text{ is decreasing.} \end{cases}$$

Therefore

$$\frac{dF_Y(y)}{dy} = \begin{cases} f_X(h(y))\,h'(y) & \text{if } g \text{ is increasing,} \\ f_X(h(y))\,(-h'(y)) & \text{if } g \text{ is decreasing,} \\ f_X(h(y))\,|h'(y)| & \text{in either case.} \end{cases}$$

Note that g is continuous everywhere, and also piecewise continuous differentiable. It follows that F_Y is the integral of its derivative:

$$F_Y(y) = \int_{-\infty}^{y} \frac{dF_Y(t)}{dt}\,dt,$$

by the fundamental theorem of calculus. Thus Y is absolutely continuous with density $f_X(h(y))\,|h'(y)|$.

Examples of this type in which g is more complicated will be considered in the problems.

Problems

1. Consider Example 4.5.6, but weaken the hypothesis on the domain of g to the union of closed intervals $I_1 \cup \cdots \cup I_n$, where on the interior of each I_j, g has a continuous nonzero derivative and is either strictly increasing or strictly decreasing, with inverse η_j. Show that g is absolutely continuous with density

$$f_Y(y) = \sum_{j=1}^{n} f_X(\eta_j(y)) \left| \eta_j'(y) \right|,$$

where $f_X(\eta_j(y)) \left| \eta_j'(y) \right|$ is interpreted as 0 if y does not belong to the domain of η_j.

2. Let X be an absolutely continuous random variable with density $f_X(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. Find the density of $Y = \min(X, 2) - \max(X, -2)$. Hint: $Y = g(X)$.

3. Let X_1, X_2, and X be independent random variables, each uniformly distributed between 0 and 1. Find the probability that $X^2 \le X_1 X_2$.

4. Let X be a random variable with density f_X, and let $Y = g(X)$, where $g : \mathbb{R}^1 \to \mathbb{R}^1$ [or $g : \mathbb{R}^n \to \mathbb{R}^n$], where n is even and $P[X \in G_j] = 1$. Assume g to be one to one and continuously differentiable with a nonzero Jacobian J_j [hence g has a continuously differentiable inverse η_j]. Show that Y is absolutely continuous with density

$$f_Y(y) = \sum f_X(\eta_j(y)) \left| J_j(\eta_j(y)) \right|^{-1}.$$

5. Let X and Y be independent random variables, each normally distributed with $\mu = 0$ and the same σ^2. Define random variables R and Θ by $R = \sqrt{X^2 + Y^2}$, $\Theta = \tan^{-1}(Y/X)$. Show that R and Θ are independent, and find their density functions (see Problem 4).

6. Let $X_1, \ldots, X_n$ be independent random variables, each with density f. Let Y_1 be the smallest of the X_j's, Y_2 the next smallest, ..., and Y_n the largest. Find $f_{Y_1, \ldots, Y_n}(y_1, \ldots, y_n)$. [Hint: Consider the event in the form $\{Y_1 \le y_1 < Y_2 \le y_2 < \cdots < Y_n \le y_n\}$ as a sum of outcomes.]

7. Let $\Omega = [0, 1]$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure. Show that $(\Omega, \mathcal{F}, P)$ is a universal probability space in the sense that if F is any proper distribution function on $\mathbb{R}^1$ [proper means that $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$], then there is a random variable X on $(\Omega, \mathcal{F}, P)$ with distribution function F. [Define $F^{-1}(y) = \sup\{x : F(x) < y\}$, $0 < y < 1$, and take $X(\omega) = F^{-1}(\omega)$, with $X(0)$ and $X(1)$ arbitrary.]

4.3.0 Introduction

Let X be a simple random variable on $(\Omega, \mathcal{F}, P)$, taking the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. If the random experiment is repeated independently N times, N very large, X will take the value x_i roughly Np_i times, so the arithmetic average of the values of X over the N trials is roughly

$$\frac{1}{N}[N_1 x_1 + \cdots + N_n x_n] \approx p_1 x_1 + \cdots + p_n x_n = \sum_{i=1}^{n} x_i p_i$$

This is a reasonable figure for the average value of X. If X is represented as $\sum_{i=1}^{n} x_i I_{A_i}$, where the A_i are disjoint sets in $\mathcal{F}$ whose union has probability one, this average value may be represented as $\sum_{i=1}^{n} x_i P(A_i)$, where $I_{A_i}(\omega) = 1$ if $\omega \in A_i$, 0 otherwise. Since arbitrary random variables are ultimately built up from simple ones, it is reasonable to take $\int_{\Omega} X \, dP$ as the definition of the average value (mathematical expectation) of X.

4.3.1 Definition. If X is a random variable on $(\Omega, \mathcal{F}, P)$, the expectation of X is defined by

$$E[X] = \int_{\Omega} X \, dP$$

provided the integral exists. Thus $E[X]$ is the integral of the Borel measurable function X with respect to the probability measure P, so that all the results of integration theory are applicable. The same definition is used if X is an extended random variable.

In many situations it is inconvenient to compute $E[X]$ by integration over Ω; the following result expresses $E[X]$ as an integral with respect to the induced probability measure P_X, which in turn is determined by the distribution function F.

4.3.2 Theorem. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with distribution function F. Let g be a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$.

If $Y = g(X)$, then

$$E[Y] = \int_{\Omega} g(X(\omega)) \, dP(\omega) = \int_{-\infty}^{\infty} g(x) \, dF(x)$$

in the sense that if either of the two integrals exists, so does the other, and the two integrals are equal.

Proof. We use the basic technique of starting with indicators and proceeding to more complicated functions.

Let g be an indicator I_A, $A \in \mathcal{B}(\mathbb{R}^n)$. Then

$$E[Y] = E[I_A(X)] = E[I_{X^{-1}(A)}] = P_X(A) = \int_A g \, dP_X$$

so that $E[Y]$ and $\int_A g \, dP_X$ exist and are equal.

Now let g be a nonnegative simple function, say $g = \sum_{i=1}^{n} a_i I_{A_i}$ for $A_i \in \mathcal{B}(\mathbb{R}^n)$ disjoint sets in $\mathcal{B}(\mathbb{R}^n)$. Then

$$E[Y] = \sum_{i=1}^{n} a_i E[I_{A_i}(X)] = \sum_{i=1}^{n} a_i \int_A I_{A_i} \, dP_X \qquad \text{by what we have just proved}$$

$$= \int_A \left(\sum_{i=1}^{n} a_i I_{A_i} \right) dP_X \qquad \text{since } a_i \geq 0$$

$$= \int_A g \, dP_X.$$

Again, both integrals exist and are equal.

If g is a nonnegative Borel measurable function, let $g_1, g_2, \ldots$ be nonnegative simple functions with $g_n \uparrow g$. We have just proved that

$$E[g_n(X)] = \int_A g_n \, dP_X$$

hence by the monotone convergence theorem,

$$E[g(X)] = \int_A g \, dP_X$$

and again both integrals exist and are equal.

Finally, let $g = g^+ - g^-$ be an arbitrary Borel measurable function and $Y = g \circ X$. We have

$$E[Y] = E[g^+(X) - g^-(X)] = E[g^+(X)] - E[g^-(X)]$$

$$= \int_A g^+ \, dP_X - \int_A g^- \, dP_X \qquad \text{by what we have already proved}$$

$$= \int_A g \, dP_X.$$

If $E[Y]$ exists and not $\infty - \infty$, then $\int_A g^+ \, dP_X$ is finite, and hence $\int_A g^- \, dP_X$ exists; by the same reasoning, the existence of $\int_A g^- \, dP_X$ implies that of $\int_A g^+ \, dP_X$. $\square$

4.20.3 *Conditional and Expectation.* (a) Let X be a random vector on $(\Omega, \mathcal{F}, P)$, and let g be a Borel measurable function such that $g \circ X$ is P-a.s. finite. Then $E(g \circ X) = \int_{\mathbb{R}^n} g \, dP_X$ in the sense that if either integral exists, so does the other, and the two are equal.

Sketch of proof: Apply 4.20.2, with $\mathbb{R}$ replaced by $\mathbb{R}^n$.

(b) More generally, let X be a random object on $(\Omega, \mathcal{F}, P)$, that is, $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, where $(\Omega', \mathcal{F}')$ is an arbitrary measurable space. Let g be a Borel measurable real (or extended real) valued function on $(\Omega', \mathcal{F}')$. For any $\mathcal{F}'$-probability measure induced by X:

$$P_X(B) = P[\{\omega : X(\omega) \in B\}] \qquad B \in \mathcal{F}',$$

then

$$E(g \circ X) = \int_{\Omega'} g \, dP_X$$

in the sense that if either integral exists, so does the other, and the two are equal.

Again the proof is just as in 4.20.2, with $\mathbb{R}$ replaced by Ω' and $\mathcal{B}(\mathbb{R})$ by $\mathcal{F}'$.

(c) If X is a random variable or random vector with density f, then

$$\int_{\Omega} g \circ X \, dP = \int_{\mathbb{R}^n} g(x) f(x) \, dx$$

in the sense that if either integral exists, so does the other, and the two are equal.

Again the proof is just as in 4.20.2, with $\mathbb{R}$ replaced by $\mathbb{R}^n$ and $\mathcal{B}(\mathbb{R})$ by $\mathcal{B}^n$.

(d) If X is a discrete random variable or random vector with probability function p, then

$$\int_{\Omega} g \circ X \, dP = \sum_{x} g(x) p(x)$$

where the series is interpreted as $\sum_{x} g^+(x) p(x) - \sum_{x} g^-(x) p(x)$, and again the interpretation is that the integral exists if the sum exists (that is,

$$\sum_{x} g^+(x) p(x) < \infty \quad \text{or} \quad \sum_{x} g^-(x) p(x) < \infty$$

and in this case the two are equal.

This is proved by starting with indicator functions as before.

Expectations of certain functions of X are so special that

4.10.4 Definition. Let X be a random variable on $(\Omega, \mathscr{S}, P)$. If $k > 0$, the number $E[X^k]$ is called the kth moment of X; $E[|X|^k]$ is called the kth absolute moment of X; $E[(X - E[X])^k]$ is called the kth central moment; $E[|X - E[X]|^k]$ is the kth absolute central moment. Central moments and absolute central moments are defined only when $E[X]$ is finite.

The first moment ($k = 1$) is $E[X]$, sometimes called the mean of X, and the first central moment (if it exists) is always 0. The second central moment $\sigma^2 = \mathrm{var}[X] = E[(X - E[X])^2]$ is called the variance of X, sometimes written $\mathrm{Var}\, X$, and the nonnegative square root σ is the standard deviation.

Note that $E[X^k]$ is finite if $E[|X|^k]$ is finite, by Exercise 4.10.3. Also, finiteness of the kth moments implies finiteness of lower moments, as we now prove.

4.10.5 Theorem. If $k > 0$ and $E[|X|^k]$ is finite, then $E[|X|^j]$ is finite for $0 < j < k$.

Proof. We have

$$E[|X|^j] = \int_{\Omega} |X|^j \, dP = \int_{|X| \le 1} |X|^j \, dP + \int_{|X| > 1} |X|^j \, dP$$

$$\le P(|X| \le 1) + \int_{\Omega} |X|^k \, dP = P(|X| \le 1) + E[|X|^k] < \infty.$$

Short cut. Prove $|X|^j \le 1 + |X|^k$ for $0 < j < k$ (Exercise 4.10.4). ■

Central moments of integer order can be obtained from moments, as follows.

4.10.6 Theorem. If n is a positive integer, and $E[|X|^n]$ is finite, and $E[X^{n-1}]$ exists, then $E[(X - E[X])^n] = \sum_{j=0}^{n} \binom{n}{j} (-1)^j E[X]^j E[X^{n-j}]$. In particular, if $E[X^2]$ is finite ($E[X]$ always exists since $X^1 \le |X|$), then

$$\mathrm{Var}\, X = E[X^2] - E[X]^2.$$

Proof. Use the binomial theorem and the additivity theorem for integrals (4.5.3). ■

A corollary formula expresses moments about the mean in terms of central moments. Write $X^n = [(X - E[X]) + E[X]]^n$ and use the binomial theorem.

We now derive a useful moment formula in a discrete-distribution context.

4.20.7 Chebyshev's inequality. Let X be a nonnegative random variable. Then for each $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof If X is a random variable with finite mean μ and variance σ^2, then for each $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

This is a quantitative result to the effect that a random variable with small variance is likely to be close to its mean.

Exercise See 2.7.5.

A normally distributed random variable has the useful property that its distribution is completely determined by the mean and variance; specifically, if X has the normal distribution, that is,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$

then $\mu = E(X)$ and $\sigma^2 = \mathrm{Var}\, X$, the computation involving the Gaussian integrals

$$\int_{-\infty}^{\infty} \exp(-x^2)\, dx = \sqrt{\pi} \qquad \text{and} \qquad \int_{-\infty}^{\infty} x^2 \exp(-x^2)\, dx = \frac{1}{2}\sqrt{\pi}.$$

The notation "$X \sim N(\mu, \sigma^2)$" is used for a random variable that is normally distributed with mean μ and variance σ^2.

The following result on the characteristic function of a product of independent random variables is a direct consequence of Fubini's theorem.

4.20.8 Theorem. Let $X_1, \ldots, X_n$ be independent random variables on $(\Omega, \mathcal{F}, P)$. If all X_j are nonnegative or all $E[X_j]$ is finite, then $E[X_1 \cdots X_n]$ exists and equals $E[X_1] \cdots E[X_n]$.

Proof If all $X_j \geq 0$, then by 4.6.3(a),

$$E[X_1 \cdots X_n] = \int_{\mathbb{R}^n} x_1 \cdots x_n \, dF_{X_1, \ldots, X_n}(x_1, \ldots, x_n) \qquad \text{where } F = (X_1, \ldots, X_n).$$

Since $f_1 \circ \pi_1$ is the product of the $f_i \circ \pi_i$, Tonelli's theorem yields

$$E[f_1(X_1) \cdots f_n(X_n)] = \int_{\mathbb{R}} f_1 \circ \pi_1 \, dP_{X_1} \cdots \int_{\mathbb{R}} f_n \circ \pi_n \, dP_{X_n} = E[f_1(X_1)] \cdots E[f_n(X_n)].$$

This can also be proved without Tonelli's theorem by starting with indicators and proceeding to nonnegative simple functions and then nonnegative measurable functions, but the present proof is clearer. Note that the result holds for extended random variables with the convention $0 \cdot \infty = 0$.

If all $E(|X_i|)$ are finite, the above argument shows that

$$E[|X_1 \cdots X_n|] = \prod_{i=1}^{n} E[|X_i|] < \infty,$$

and Fubini's theorem may be applied just as in the first part of the proof. □

Theorem 4.10.8 can be extended to complex valued random variables. To see this, recall that a complex valued random variable λ on $(\Omega, \mathcal{F}, P)$ is given by $\lambda = X_1 + iX_2$ where $X_1 = \operatorname{Re} \lambda$ and $X_2 = \operatorname{Im} \lambda$ are real-valued random variables. In view of the theorem at the beginning of 4.7 we may regard λ as simply a two-dimensional random vector. Moreover $\lambda(\omega) = \lambda(\omega')$ provided $X_1(\omega) = X_1(\omega')$ and $X_2(\omega) = X_2(\omega')$.

4.10.9 Theorem. *Let $\lambda_1, \ldots, \lambda_n$ be independent complex valued random variables and $E(|\lambda_i|)$ is finite for all i. Then $E[\lambda_1 \cdots \lambda_n]$ is finite and equals $E[\lambda_1] \cdots E[\lambda_n]$.*

Proof. For $n = 2$, $\lambda_1 = X_1 + iX_2$, $\lambda_2 = Y_1 + iY_2$. By 4.10.8 X_1, X_2, Y_1 and Y_2 are independent, as are X_1 and Z_2, Z_1 and X_2, and Z_1 and Z_2. By 4.10.8,

$$E[\lambda_1 \lambda_2] = E[(X_1 + iX_2)(Y_1 + iY_2)] = E[X_1 Y_1] - E[X_2 Y_2] + iE[X_1 Y_2] + iE[X_2 Y_1]$$

$$= E[X_1] E[Y_1] - E[X_2] E[Y_2] + iE[X_1] E[Y_2] + iE[X_2] E[Y_1] = E[\lambda_1] E[\lambda_2].$$

Now let $n > 2$, $\lambda_1 \cdots \lambda_n = (X_1 + iX_2)(Y_1 + iY_2)$, and we assume the result has been established for $n - 1$ complex variables. Let $V = (X_1, Z_1, X_2, Z_2, \ldots, X_{n-1}, Z_{n-1})$ and $W = (Z_n, Z_{n+1})$. We claim that V and W are independent. By independence of $X_1, \ldots, X_n$, $P[V \in A, W \in B] = P[V \in A] P[W \in B]$ when A and B are cylinder sets. Thus from measurable rectangles to finite disjoint unions of measurable rectangles and then, by means of the monotone class theorem, to arbitrary Borel sets.

The independence of V and W implies, by 4.10.8, that $X_1 \cdots X_{n-1} Y_1$ and Y_2 are independent, so that $E[X_1 \cdots X_{n-1} Y_1 Y_2] = E[X_1 \cdots X_{n-1} Y_1] E[Y_2] = E[X_1] \cdots E[Y_1] E[Y_2]$ by the inductive hypothesis. □

Theorem 4.10.8 implies that the variance of a sum of independent random variables is the sum of the variances. Actually, zero correlation is all that is needed. We now introduce some new terminology.

4.10.10 **Definitions and Comments.** Let X and Y be random variables with finite expectations, and assume $E(XY)$ is also finite. (In particular, if X and Y have finite second moments, then XY is integrable by the Cauchy–Schwarz inequality.) The covariance of X and Y is defined by

$$\mathrm{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ by 4.10.5; however, the converse is not true (consider $X = \sin\theta$, $Y = \cos\theta$, where θ is uniformly distributed between 0 and 2π).

If the variances σ_X^2 and σ_Y^2 are finite and greater than 0, the correlation coefficient between X and Y is defined by

$$\rho(X, Y) = \rho[X, Y] = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

By the Cauchy–Schwarz inequality applied to $X - E(X)$ and $Y - E(Y)$, $-1 \le \rho(X, Y) \le 1$. Equality occurs (see Problem 6, Section 2.4) if $X - E(X) = c(Y - E(Y))$ and $Y - E(Y) = d(X - E(X))$ are linearly dependent (that is, $P[aX + bY = k] = 1$ for some real numbers a, b, k with a and b not both 0).

We now look at the variance of a sum.

4.10.11 **Theorem.**

If $X_1, \ldots, X_n$ are random variables with finite expectations and $E(X_i X_j) < \infty$ for all i, j with $i \neq j$,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}\, X_i + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Thus if $Y_1, \ldots, Y_n$ are pairwise uncorrelated, that is, $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$, in particular, if $X_1, \ldots, X_n$ are independent, then

$$\mathrm{Var}(Y_1 + \cdots + Y_n) = \sum_{i=1}^{n} \mathrm{Var}\, Y_i.$$

Proof.

$$\begin{aligned}
&\mathrm{Var}(X_1 + \cdots + X_n) \\
&= E[(X_1 + \cdots + X_n)^2] - [E(X_1) + \cdots + E(X_n)]^2 \\
&= \sum_{i=1}^{n} \{E[X_i^2] - [E(X_i)]^2\} + 2 \sum_{i<j} \{E[X_i X_j] - E(X_i)E(X_j)\}
\end{aligned}$$

4.10.12 *Corollary.* Under the hypothesis of 4.10.11, if $a_1, \ldots, a_n$ are real numbers then

$$\operatorname{Var}(a_1 X_1 + \cdots + a_n X_n) = \sum_{i=1}^{n} a_i^2 \operatorname{Var} X_i + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j).$$

Proof. This follows from 4.10.11, along with the observations which may be recalled from the definitions, that $\operatorname{Var}(aX) = a^2 \operatorname{Var} X$ and $\operatorname{Cov}(X_i, a_j X_j) = a_j \operatorname{Cov}(X_i, X_j)$. ∎

Problems

1. Let X be a discrete random variable with $P(X = n) = \frac{1}{2^n}$, $n = 1, 2, \ldots$. Let $Y = g \circ X$, where $g(n) = (-1)^n 2^n / n$. Show that $E(Y)$ does not exist. (Compare the series $\sum_{n=1}^{\infty} n P(X = n)$ is conditionally convergent.)

2. Let X be a random variable with the distribution function shown in Figure 4.10.1. Compute $E(X^2)$.

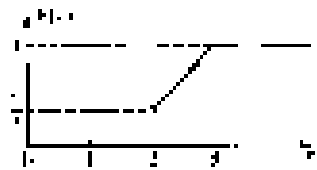

Figure 4.10.1

3. Suppose X is a random variable with distribution function F_X, and $Y = g \circ X$, $g : \mathbb{R} \to \mathbb{R}$ Borel measurable. It is desired to find the expectation of Y. One student evaluates $\int_{-\infty}^{\infty} g(x) dF_X(x)$; another first finds the distribution function of Y, that is,

$$F_Y(y) = P[Y \le y] = P[X \in g^{-1}(-\infty, y]]$$

and then evaluates $\int_{-\infty}^{\infty} y \, dF_Y(y)$. Are the answers the same?

4. Show that n random variables $X_1, \ldots, X_n$ are independent iff for all Borel measurable $g : \mathbb{R} \to \mathbb{R}$ such that $g_i \ge 0$, we have

$$E[g_1(X_1) \cdots g_n(X_n)] = \prod_{i=1}^{n} E[g_i(X_i)] \quad \text{where} \quad g_i(X_i) = g_i \circ X_i. \tag{1}$$

Also show that $X_1, \ldots, X_n$ are independent iff (1) holds for all Borel measurable $g : \mathbb{R} \to \mathbb{R}$ such that $E[g_i(X_i)]$ is finite for each i.

5. Let X be an extended random variable on $(\Omega, \mathcal{F}, P)$, and let u be a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$. Define $F(x) = P[u \circ X \le x]$, $x \in \mathbb{R}$. Define $F(x-) = \lim_{y \uparrow x} F(y)$. Show that if $E[u \circ X] < \infty$, then $E[u \circ X] = \int_{-\infty}^{\infty} u(x)\, dF(x)$. (Hint: $P[X = \infty] = 1 - F(\infty-)$.)

4.11 Infinite Sequences of Random Variables

Very often we will be interested in a sequence of random variables $X_1, X_2, \ldots, X_n$, where n is arbitrarily large and not fixed in advance. This is conveniently viewed as a single probability space on which we can define an infinite sequence of random variables. The discrete product measures on infinite-product spaces (2.7) provide the necessary machinery. In particular, we can assume that $X_1, X_2, \ldots$ are independent with X_j having a specified distribution. In fact the X_j can be random objects with values in an arbitrary measurable space.

4.11.1 *Theorem.* Let $(\Omega_j, \mathcal{F}_j, P_j)$, $j = 1, 2, \ldots$, be an arbitrary sequence of probability spaces. There exists a probability space $(\Omega, \mathcal{F}, P)$ and a sequence of independent random objects $X_j : (\Omega, \mathcal{F}) \to (\Omega_j, \mathcal{F}_j)$ such that

$$P[X_j \in A] = P_j(A) \qquad \text{for all } A \in \mathcal{F}_j,\ j = 1, 2, \ldots.$$

To obtain a sequence of independent random variables with specified distributions, for instance $N(0,1)$, we take $\Omega_j = \mathbb{R}$, $\mathcal{F}_j = \mathcal{B}(\mathbb{R})$, with P_j the Lebesgue–Stieltjes measure corresponding to F_j.

Proof. Let $\Omega = \prod_{j=1}^{\infty} \Omega_j$, $\mathcal{F} = \prod_{j=1}^{\infty} \mathcal{F}_j$, $P = \prod_{j=1}^{\infty} P_j$ (see 2.7.7). If $\omega = (\omega_1, \omega_2, \ldots) \in \Omega$, set $X_j(\omega) = \omega_j$, $j = 1, 2, \ldots$. If $A \in \mathcal{F}_j$, then $[X_j(\omega) \in A] = [\omega : \omega_j \in A]$ is a measurable rectangle. Thus for $A \in \mathcal{F}_j$, $P[X_j \in A] = P_j(A)$, so that the X_j are random objects with $X_j : (\Omega, \mathcal{F}) \to (\Omega_j, \mathcal{F}_j)$.

$$P[X_{j_1} \in A_1, \ldots, X_{j_n} \in A_n] = \prod_{i=1}^{n} P_{j_i}(A_i) \qquad \text{if } A_1 \in \mathcal{F}_{j_1}, \ldots, A_n \in \mathcal{F}_{j_n}.$$

Take $A_i = \Omega_{j_i}$ for $i \ne k$. It follows that $P[X_{j_k} \in A_k] = P_{j_k}(A_k)$. Therefore $P[X_{j_1} \in A_1, \ldots, X_{j_n} \in A_n] = \prod_{i=1}^{n} P[X_{j_i} \in A_i]$. This implies independence. $\square$

The results of (1) may also be used to describe the underlying probability space for a Markov chain. Let S be a finite or countably infinite set, called the state space, for convenience assumed also $\mathcal{F}$-measurable. Let $\Pi = [p_{ij}]_{i,j \in S}$ be a transition matrix, that is, $p_{ij} \ge 0$ for all $i, j \in S$ and $\sum_{j \in S} p_{ij} = 1$ for all $i \in S$. Let μ_i, $i \in S$, be a set of nonnegative numbers adding to 1. The Markov chain corresponding to Π and $\{\mu_i\}$ is a process that starts at time $n = 0$ in state i with probability μ_i, where $P[X_0 = i] = \mu_i$, $i \in S$, and makes transitions at

times $t = 0, 1, \ldots$ in accordance with the following rule: If X_n denotes the state at time n, and we know that $X_n = i$, then regardless of the past history of the process, the probability of moving to state j at time $n+1$ is p_{ij}. We expect from (7.1.5) that for all $i_0, i_1, \ldots, i_{n-1}, i, j \in S$, $n = 0, 1, \ldots$,

$$P\{X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = p_{ij}. \tag{7.1.6}$$

We now show that it is possible to construct a sequence of random variables satisfying this requirement.

7.1.2 Theorem *Given a finite or countable set S, a stochastic matrix $\Pi = [p_{ij}]$, $i, j \in S$, and a distribution $\{p_i : i \in S\}$, there is a sequence of random variables $X_0, X_1, \ldots$, all defined on the same probability space and taking values in S, such that*

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\} = p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n} \tag{7.1.7}$$

for all $i_0, \ldots, i_n \in S$ and all $n = 0, 1, \ldots$.

Proof Let $\Omega = S^{\infty}$, the set of all sequences $\omega = (\omega_0, \omega_1, \ldots)$ with $\omega_n \in S$. Define

$$P_0(B_0) = \sum_{i_0 \in B_0} p_{i_0} \qquad B_0 \subset S$$

$$P\{\omega_0 = i_0, \ldots, \omega_n = i_n\} = p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n} \qquad i_0, \ldots, i_n \in S.$$

Since $\mathcal{F}^1$ consists of all subsets of S^1, the measurability requirements of Theorem 7.1.2 are automatically satisfied. If we define $X_n(\omega_0, \omega_1, \ldots) = \omega_n$, $n = 0, 1, \ldots$, we obtain

$$P\{X_0 = i_0, \ldots, X_n = i_n\} = \int_{S_0} P_0(d\omega_0) \int_{S_1} P(\omega_0, d\omega_1) \cdots \int_{S_n} P(\omega_{n-1}, d\omega_n),$$

where $S_k = \{\omega_k : (\omega_0, \ldots, \omega_k) \in A\}$, $A = \{\omega : \omega_0 = i_0, \ldots, \omega_n = i_n\}$. Now if μ is a measure on S^1 with $\mu\{i\} = q_i$, $i \in S$, and P is a transition probability function, then $\int_S P(\omega, B)\,\mu(d\omega) = \sum_{i \in S} q_i P(i, B)$; see (6.1.19). Thus

$$\int_{S_{n-1}} P(\omega_{n-1}, S_n)\, P\{\omega_0 = i_0, \ldots, \omega_{n-1} = i_{n-1}, d\omega_{n-1}\}$$

$$= P\{\omega_0 = i_0, \ldots, \omega_{n-1} = i_{n-1}\} p_{i_{n-1} i_n}$$

$$= p_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

We may generate this process to obtain

$$P(X_{n+1} = i_0, \ldots, X_{n+k} = i_k \mid X_n = i_0) = p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{k-1} i_k}.$$

A sequence of random variables $\{X_n\}$ satisfying (4.1.1) is called a *Markov chain* corresponding to the matrix Π and the initial distribution $\{p_i\}$. Π is called the transition matrix of the chain and the numbers p_{ij} are called *transition probabilities*.

The basic properties of Markov chains are discussed in Feller (1970, Chapters XV and XVI). The symmetric random walk on $\mathbb{Z}^d$, an important special case, is considered in Appendix I.

If $X_1, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F}, P)$ for certain sets $X = \{X_1, \ldots, X_n\}$ is a finite measurable map from Ω to $\mathbb{R}^n$. The same interpretation is possible for an infinite sequence of random variables as follows.

4.1.2 *Theorem.* Let $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2, \ldots$, be arbitrary measurable spaces and let p_n be the projection of $\prod_{i=1}^{\infty} \Omega_i$ onto the ith coordinate space. If $(\Omega, \mathcal{F})$ is a measurable space, $X: \Omega \to \prod_{i=1}^{\infty} \Omega_i$, let $X_i = p_i \circ X$, $i = 1, 2, \ldots$. Then X is measurable relative to $\mathcal{F}$ and $\prod_{i=1}^{\infty} \mathcal{F}_i$ if and only if each X_i is measurable relative to $\mathcal{F}$ and $\mathcal{F}_i$. In particular, if $X_1, X_2, \ldots$ are random variables, then $X = (X_1, X_2, \ldots)$ is measurable relative to $\mathcal{F}$ and $\mathcal{B}^{\infty} = \mathcal{B} \times \mathcal{B} \times \cdots$. For this reason X is sometimes called a random sequence. The above result tells us that Ω is a relatively (possibly uncountably) infinite space.

Proof. If X is measurable, each X_i is a composition of measurable maps and is therefore measurable. Conversely, assume each X_i is measurable. Let $\mathcal{A} = \{\prod_{i=1}^{\infty} A_i : A_i \in \mathcal{F}_i$ for all i, and $A_i = \Omega_i$ for all but a finite number of $i\}$. It is easy to check that $\mathcal{A}$ is an algebra and $\sigma(\mathcal{A}) = \prod_{i=1}^{\infty} \mathcal{F}_i$. Thus

$$\{\omega \in \Omega : X(\omega) \in B\} = \bigcup_{\text{finite}}^{\infty} \{\omega \in \Omega : X_i(\omega) \in A_i\} \in \mathcal{F}.$$

The proof for uncountable product spaces is essentially the same, with $\mathbb{N}$ replaced by a measurable rectangle in $\prod_{\alpha} \mathcal{F}_{\alpha}$. $\square$

The proof of 4.1.2 shows also that if $X: \Omega \to \prod_{i=1}^{\infty} \Omega_i$, that X is measurable iff each X_i, $1 \le i \le n$, is measurable.

We close this chapter with an introduction to one of the basic limit theorems of probability.

4.1.3 *Weak Law of Large Numbers.* Let $X_1, X_2, \ldots$ be independent random variables (not necessarily with the same distribution), each with finite mean

and variance. Assume the variances to be uniformly bounded by M, i.e., $\mathrm{Var}\,X_n \le M$ for $n = 1, 2, \ldots$. Then $\frac{S_n}{n} - \frac{E(S_n)}{n}$ converges in probability to 0, i.e., for any $\varepsilon > 0$,

$$P\left\{\left|\frac{S_n}{n} - \frac{E(S_n)}{n}\right| \ge \varepsilon\right\} \to 0 \quad \text{as} \quad n \to \infty.$$

Proof. By Chebyshev's inequality,

$$P\{|S_n - E(S_n)| \ge n\varepsilon\} \le \frac{1}{n^2\varepsilon^2} E\left|\left(S_n - E(S_n)\right)^2\right|$$

$$= \frac{1}{n^2\varepsilon^2} \mathrm{Var}\,S_n$$

$$= \frac{1}{n^2\varepsilon^2} \sum_{i=1}^{n} \mathrm{Var}\,X_i \qquad \text{by 4.10.1}$$

$$\le \frac{M}{n\varepsilon^2} \to 0.$$

There are two special cases of particular interest.

1. If $P\{X_i = a\} = 1$ for all i, then $\frac{S_n}{n} - E\left(\frac{S_n}{n}\right) = \frac{S_n}{n} - a$ converges to 0 in probability. Thus for large n, the arithmetic average of n independent random variables each with finite expectation a and with the variances uniformly bounded is quite likely to be very close to a.

2. If $X_1, X_2, \ldots$ are independent and for each i, $P\{X_i = 1\} = p$, $P\{X_i = 0\} = q = 1 - p$, then we have an infinite sequence of Bernoulli trials, and $X_1 + \cdots + X_n$ is the number of successes in n trials, hence $\frac{S_n}{n}$ is the relative frequency of successes. Since $E(X_i) = p$, we have $\frac{S_n}{n} \to p$ in probability. Thus for large n, the relative frequency of success is quite likely to be very close to p.

Informally, the weak law of large numbers says the following. If an original observation of $X_1, \ldots, X_n$ is one performance of an experiment, where n is very large but fixed, then if we repeat the experiment independently, $\frac{S_n}{n} - E\left(\frac{S_n}{n}\right)$ will be close to 0 in a very high percentage of the trials.

More physically we expect something more than this. If a coin with probability p of heads is tossed over and over again, we expect the relative frequency to approach p in the ordinary sense of convergence of a sequence of real numbers; in other words, given $\varepsilon > 0$, however small, we expect the relative frequency $\frac{S_n}{n}$ to get and remain within ε of p. Here we are considering a realization of the infinite sequence $X_1, X_2, \ldots$, one performance of the experiment, and what we mean

show in that $\lim_{n\to\infty} X_n(\omega) = \rho$ for almost every ω. A statement of this type is called a strong law of large numbers. This subject will be considered in more detail later.

Problems

1 (a) If $X_1, X_2, \ldots$ are independent random variables, show that for each n the variables $(X_1, \ldots, X_n)$ and $(X_{n+1}, X_{n+2}, \ldots)$ are independent.

(b) If part (a) of $\{X_n\}$ $(\Omega, \mathcal{F}, P)$ $\omega \in \Omega$ for all n, and all X_n have the same distribution, show that $(X_1, X_2, \ldots)$ and $(X_n, X_{n+1}, \ldots)$ have the same distribution for all n.

2 Consider the gambler's ruin problem, that is the simple random walk with absorbing barriers at 0 and b as in the problems. $X_1, X_2, \ldots$ are independent random variables, with $P\{X_i = 1\} = p$, $P\{X_i = -1\} = q = 1 - p$. Let $Y_n = \sum_{i=1}^{n} X_i$, and let a be an arbitrary integer between 1 and $b - 1$. We wish to find the probability $h(a)$ of eventual ruin starting from a, in other words the probability that $a + Y_n$ will reach 0 before it reaches b. Intuitive reasoning, based on the fact that the total probability leads to the result that

$$h(a) = ph(a+1) + qh(a-1) \quad \text{for } a = 1, \ldots, b-1.$$

Give a formal proof of this result. (For further details see Ash, 1970, Chapter 6.)

4.12 Bibliography

The present treatment of this chapter is based on Ash (1970b), which is an excellent introductory course in probability. Measure theory is not used in the book, although some of the underlying measure theoretic ideas are discussed. Many additional examples and problems can be found in Feller (1968) and Parzen (1960).

To develop intuitive skills we recommend *One Thousand Exercises in Probability* by G. Ash (1970). Another useful reference is Rao (1973).

CHAPTER

5

CONDITIONAL PROBABILITY AND EXPECTATION

5.1 INTRODUCTION

In Chapter 4, we defined the conditional probability $P(F|E)$ only when $P(E) > 0$. However, conditional probabilities given events of probability zero can sometimes be defined, by using limiting arguments. For example, suppose that for every positive integer n, an experiment is performed whose outcome is a random variable X_n whose distribution function F_n is continuous. If X_n does take on the value x, a random variable Y_n is observed, whose distribution is $F_{n,x}$. For example, if X_n is uniform on $(0, n)$, then Y_n might be the number of independent tosses of a coin, with P(heads) $= p$, that are necessary to obtain n heads, where $p = x/n$. The conditional distribution $P\{Y_n \le y | X_n = x\}$ is prescribed in the statement of the problem, although the event $X_n = x$ may have probability zero for all values of x.

The reader is invited to consider the above situation. Let $F_n(x, y) = P\{X_n \le x, Y_n \le y\}$ denote the joint distribution function of X_n and Y_n. We specify the distribution function of X_n and the collection of corresponding probability measures $P_{n,x}(\cdot)$, where for each x and n given a probability measure $P_{n,x}(\cdot)$ defined on $(-\infty, \infty)$. $P_{n,x}(B)$ is interpreted (informally for now) as $P\{Y_n \in B | X_n = x\}$.

We claim that the probability of any event of the form $\{X_n \in A, Y_n \in B\}$ is determined. Reasoning informally, the probability that X_n is in $(x, x + dx)$ and Y_n is in B is $P_{n,x}(B)\,dF_n(x)$. Given that this occurs, in other words, roughly, given $X_n = x$, Y_n, $P\{Y_n \in B\}$ is $P_{n,x}(B)$ if Y_n belongs to B and is 0 if Y_n does not belong to B. The probability of this event is $P_{n,x}(B)\,dF_n(x)$. The total probability that X_n will belong to A is

$$P\{X_n \in A\} = \int_{-\infty}^{\infty} P_{n,x}(A)\,dF_n(x) \tag{5.1}$$

In the special case $S = \{x : x \in A, x \in B\} = A \cap B$, $C(x) = 0$ if $x \in A \cap B$ and $C(x) = 1$ if $x \in A$. Therefore

$$P(S) = P(A \cap B) = \int_A P(x, B)\, dP(x). \tag{2.}$$

Now if $P(x, B)$ is Borel measurable in x for each fixed $B \in \mathcal{B}(\mathcal{X}_2)$, then by the product measure theorem, there is a unique (probability) measure on $\mathcal{B}(\mathcal{X}_1) \times \mathcal{B}(\mathcal{X}_2)$ satisfying (2) for all $A \in \mathcal{B}(\mathcal{X}_1)$, namely, the measure given by (1). Given the marginal measure P on $\mathcal{X}_1$ and the transition probability measure $P(x, \cdot)$, $x \in \mathcal{X}_1$, we take the probability measure P on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ to be the unique measure determined by (1) and the measure on $\mathcal{X}_1 \times \mathcal{X}_2$ is denoted by $P(x, \cdot)\, dP(x)$.

3.2 Applications

We apply the results of 3.1 to some typical situations in probability.

3.2.1 Example. Let X be uniformly distributed between 0 and 1. If $X = x$, a coin with probability x of heads is tossed independently a number of times. If T is the waiting number of heads, find $P(T = k)$, $k = 0, 1, \ldots$.

Let us translate this information into mathematical terms. Let $\Omega_1 = [0, 1]$, $\mathcal{F}_1 = \mathcal{B}[0, 1]$. We have specified $P_1(A) = \int_A dx$ = Lebesgue measure of A, $A \in \mathcal{F}_1$.

For each x, we are given $P(x, B)$ for $x \in \Omega_1$, B in the σ-field of all subsets of Ω_2 that $P(x, \{k\}) = \binom{n}{k} x^k (1 - x)^{n-k}$, $k = 0, 1, \ldots, n$. This is Borel measurable in x. We take $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{B}_1 \times \mathcal{B}_2$, P the unique probability measure determined by P_1 and the $P(x, \cdot)$, namely

$$P(\Gamma) = \int_0^1 P(x, \Gamma(x))\, dP_1(x) = \int_0^1 P(x, \Gamma(x))\, dx.$$

Now let $T(x, y) = x$, $T(x, y) = y$. Then

$$P(T = k) = P(\Omega_1 \times \{k\}) = \int_0^1 P(x, \{k\})\, dx$$

$$= \int_0^1 \binom{n}{k} x^k (1 - x)^{n-k}\, dx = \binom{n}{k} \beta(k + 1, n - k + 1)$$

where $\beta(r, s) = \int_0^1 x^{r-1} (1 - x)^{s-1}\, dx$, $r, s > 0$, is the beta function. We can express $\beta(r, s)$ as $\Gamma(r)\Gamma(s)/\Gamma(r + s)$, where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx$, $t > 0$, is the gamma function. Since $\Gamma(n + 1) = n!$, $n = 0, 1, \ldots$, we have

$$P(T = k) = \frac{n!}{k!(n - k)!} \frac{k!(n - k)!}{(n + 1)!} = \frac{1}{n + 1}, \qquad k = 0, 1, \ldots, n.$$

In solving a problem of this type, what we have in mind is a real or [illegible]

$x_1, \ldots, x_n$ when we know that $\Lambda = \lambda$, the probability that X heads will be obtained is $\binom{n}{x}\lambda^x(1-\lambda)^{n-x}$. Integrating this function with respect to the total probability measure of Λ gives the unconditional probability of x successes in n trials in the random experiment.

5.2.2 *Example.* Let Z_1 be uniformly distributed between 0 and 1. If $Z_1 = z_1$, let Z_2 be uniformly distributed between 0 and z_1. In general, if $Z_1 = z_1, \ldots, Z_{n-1} = z_{n-1}$, let Z_n be uniformly distributed between 0 and z_{n-1}. For $n = 1, 2, \ldots$ find the expectation of Z_n.

Here we have $\Omega_n = \mathbb{R}$, $\mathcal{B}_n = \mathcal{B}(\mathbb{R})$. $\Omega = \prod_{n=1}^{\infty} \Omega_n$, $\mathcal{B} = \prod_{n=1}^{\infty} \mathcal{B}_n$. $X_n(\omega) = \omega_n$ if $\omega = (\omega_1, \omega_2, \ldots) \in \Omega$, $n = 1, 2, \ldots$.

Let μ = Lebesgue measure on $(0, 1)$, and for each $x \in (0, 1)$,

$$P(x, \cdot) = \frac{1}{x}[\text{Lebesgue measure on } (0, x)],$$

that is,

$$P(x, B) = \frac{1}{x}\int_{B \cap (0,x)} dz.$$

For general, for each $x_1, \ldots, x_n \in (0, 1)$, $B \in \mathcal{B}_{n+1}$, $n = 1, 2, \ldots$, take

$$P(x_1, \ldots, x_n, \cdot) = \frac{1}{x_n}[\text{Lebesgue measure on } (0, x_n)]$$

(We must show in general, as we did for $n = 1$.)

Let P be the unique measure on $\mathcal{B}$ determined by P and the $P(x_1, \ldots, x_n, \cdot)$. We may find the expectation of a Borel measurable function u from $\mathbb{R}^n$ to $\mathbb{R}$ by Tonelli's Theorem:

$$\int_{\Omega} u\, dP = \int_0^1 P(1, x_1)\, dx_1 \int_0^1 P(x_1, x_2)\, dx_2 \cdots \int_0^1 g(x_1, \ldots, x_n)\, dx_n$$
$$\times P(x_1, \ldots, x_n, \Omega_{n+1}).$$

In the present case we have $g(x_1, \ldots, x_n) = x_1 x_2 \cdots x_n$. Thus

$$E(Z_n) = \int_0^1 dx_1 \int_0^{x_1} \frac{1}{x_1} dx_2 \int_0^{x_2} \frac{1}{x_2} dx_3 \cdots \int_0^{x_{n-1}} \frac{x_n}{x_{n-1}} dx_n$$
$$= \int_0^1 \frac{x_1}{2^{n-1}}\, dx_1 = 2^{-n}.$$

This example has an alternative interpretation. Let $Y_1, \ldots, Y_n$ be independent random variables each uniformly distributed between 0 and 1. Let Z_n be the greatest Y_k less than or equal to the others (for example, if $n = 3$, $Y_1 = .3$, $Y_2 = .5$, $Y_3 = .4$, then $Z_3 = .4$). Show that $(Z_1, \ldots, Z_n)$ has the same distribution as $(\Omega_1, \ldots, \Omega_n)$. Deduce $E(Z_n)$.

$= E(Z_1)$. Since the expectation of a product of independent nonnegative random variables is the product of the expectations $E(Z_n) = [E(Z_1)]^n = 2^n$ as before.

5.2.4 Example. Let X be a discrete random variable taking on the positive integer values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, $p(x_i) = p_i > 0$, $\sum_{i=1}^{\infty} p_i = 1$. If $X = x_i$, a nonnegative number X is selected according to the density f_i. Find the probability that $1 \le X \le 2$.

Here we have

$$\Omega = \{1, 2, \ldots\} \qquad \mathcal{F} = \text{the class of all subsets of } \Omega$$

$$P(\{i\}) = p_i \qquad i = 1, 2, \ldots$$

$$\Omega_i = R_1 \qquad \mathcal{B}_i = \mathcal{B}(R_1) \qquad P_i(B) = \int_B f_i(x)\,dx$$

$$X(\omega_1, \omega_2) = \omega_1 \qquad Y(\omega_1, \omega_2) = \omega_2.$$

The measure determined by P^* and the $P(\omega_1)$ is given by

$$P^*(F) = \int_{\Omega_1} P(\omega_1) P_{\omega_1}(F^{\omega_1})(\omega_1)\,dP(\omega_1) \qquad F \in \mathcal{F}^*$$

$$= \sum_{i=1}^{\infty} P(\{i\}) P_i(F^i)(\omega_i) \qquad \omega_i \in \{1, 2, \ldots\}.$$

If $F = \{(\omega_1, \omega_2) : \omega_1 = \omega_2 \in B\}$, then $F^i(\omega_i) = \{\omega_2 : (i, \omega_2) \in F\} = B$, and hence

$$P^*(F) = P[1 \le X = Y \le 2] = \sum_{i=1}^{\infty} p_i \int_1^2 f_i(x)\,dx$$

$$= p_1 \int_1^2 f_1(x)\,dx + p_2 \int_1^2 f_2(x)\,dx \quad \text{since } f_i(x) = 0 \text{ for } i > 2.$$

Note that if $p_1 = 0$ then $P[Y = 2|X = 1]$ is undefined and only $P[Y = n, X = 1] = P[Y = n | X = 1] P[X = 1] = 0$. But $P[X = 1, Y = n] = P[\{(\omega_1, \omega_2) : \omega_1 = 1, \omega_2 = n\}] = P(\{1\}, B) = p_1 P_1(B)$. Thus $P[Y \in B | X = 1] = P_1(B)$, as we would expect intuitively.

For additional examples see Ash (1970, Chapter 4).

5.3 The General Concept of Conditional Probability and Expectation

We have seen that specification of the distribution of a random variable Y together with $f_{X|Y}(x, y)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, interpreted intuitively as the conditional probability that $X \in B$ given $Y = y$, determines a unique probability

measure on $\mathcal{F}(\mathcal{X})$ on the line. There is only one reasonable joint distribution of X and Y consistent with the given data. However, the symmetrical elegant approach to the existence and uniqueness of regular conditional probabilities given events with probability zero (as in Examples 5A to 5C) is not just on $(\Omega, \mathcal{A}, P)$, but in $(\Omega, \mathcal{A}, P) \times (\Omega', \mathcal{A}')$ and $B \in \mathcal{A}'$, we may ask whether there is a useful definition of a meaningful way to find conditional probability $P(x, B) = P(Y \in B | X = x)$, even though the event $(X = x)$ may have probability zero for some, in fact perhaps for all, x.

By the discussion in 5.1, if we have a reasonable conditional probability $P(x, B)$, it should satisfy

$$P(X \in A, Y \in B) = \int_A P(x, B)\, dP_X(x),$$

where P_X is the probability measure induced by X, namely

$$P_X(A) = P(\{\omega : X(\omega) \in A\}), \quad A \in \mathcal{A}.$$

In fact, this requirement determines $P(x, B)$ in the following sense.

5.3.1 **Theorem** Let $X : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ be a random variable on $(\Omega, \mathcal{A}, P)$, and let B be a fixed set in $\mathcal{A}'$. Then there is a unique (Borel) measurable function g on $(\Omega', \mathcal{A}')$ such that for each $A \in \mathcal{A}'$,

$$P(\{Y \in B\} \cap \{X \in A\}) = \int_A g(x)\, dP_X(x).$$

Furthermore, if h is another such function then $g = h$ a.e. $[P_X]$. [We define $P(B | X = x)$ as $g(x)$. It is essentially unique for a.e. $x[P_X]$.]

Proof Let $\nu(A) = P(\{Y \in B\} \cap A)$, $A \in \mathcal{A}$. Then ν is a finite measure on $\mathcal{A}$, absolutely continuous with respect to P. ($P(A) = 0$ implies $\nu(A) = 0$.) The result follows from the Radon–Nikodym theorem. □

Let us now verify that the conditional probability we have just constructed coincides with our intuition in simple cases.

5.3.2 **Examples** (a) Let X take on only countably many values $x_1, x_2, \ldots$ with

$$p_i = P(X = x_i) > 0, \qquad \sum_{i=1}^{\infty} p_i = 1.$$

We claim that

$$g(x_i) = P(B | X = x_i) = \frac{P(B \cap \{X = x_i\})}{P(X = x_i)}, \qquad i = 1, 2, \ldots$$

([illegible] is constructed on the σ_n we need only be able to specify sets for which integrals exist.) This is a general definition reducing to the elementary definition in the discrete case. To prove this, let $\mathcal{C}' = \{\sigma_1, \ldots, \sigma_k\}$ with $\mathcal{C}'$ the collection of all subsets of $\mathcal{C}'$. If $A \in \mathcal{C}'$ and σ is defined as above, then

$$\int_A g(x)\,dP_\sigma(x) = \int_{\sigma^{-1}(A)} g(\sigma(t))\,dP(t) = \sum_{x \in A} g(x)\,P(\sigma^{-1}(x)) \quad \text{[see 4.1.3 (b)(c)]}$$

$$= \sum_{x \in A} g(x)\,P[X = x] = \sum_{x \in A} P(B^{-1}(\{x\}))$$

$$= P_\sigma[X \in A] = P_\sigma(A).$$

Since there is essentially only one σ satisfying

$$\int_{\sigma^{-1}(A)} g(\sigma)\,dP = P[\{X \in A\}]\,P(\sigma^{-1}(A)), \quad A \in \mathcal{C}',$$

this is the proposal must be correct.

Example. Let X and Y be random variables with joint density f

$$[a, b] \times [c, d], \quad \mathcal{P} = \mathcal{B}[a, b] \times \mathcal{B}[c, d], \quad \sigma(x, y) = x, \quad P(x, y) = y,$$

$$P(A) = \int_A \int_c^d f(x, y)\,dy\,dx \qquad A \in \mathcal{B}[a, b].$$

Now $P[X = x]$ is probability zero for each x, but there is a reasonable approach to the conditional probability $P[Y \in C \mid X = x]$, as follows:

$$P_{\sigma^{-1}(\mathcal{C}')}(Y \in C)(x) = P[Y \in C \mid X \in A_\varepsilon] = \frac{P[Y \in C,\ X \in A_\varepsilon]}{P[X \in A_\varepsilon]}$$

$$= \frac{\int_{A_\varepsilon} \int_C f(x, y)\,dy\,dx}{\int_{A_\varepsilon} f_1(x)\,dx},$$

where $f_1(x) = \int_c^d f(x, y)\,dy$ is the density of X.

For small ε, this is (for each y) approximately

$$\frac{2\varepsilon \int_C f(x, y)\,dy}{2\varepsilon f_1(x)} = \int_C \frac{f(x, y)}{f_1(x)}\,dy$$

We are led to define

$$h(y \mid x) = \frac{f(x, y)}{f_1(x)}$$

as the conditional density of Y given $X = x$ for those x in the support of f_1, the marginal density of X given X. Note that h is defined only when $f_1(x) \neq 0$. However, if $S = \{x : f_1(x) = 0\}$, then $P[(X, Y) \in S] = 0$, since

$$P[(X, Y) \in S] = \int_S \int f(x, y)\, dy\, dx = \int_{(x : f_1(x) = 0)} \left| \int_{-\infty}^{\infty} f(x, y)\, dy \right| dx$$

$$= \int_{(x : f_1(x) = 0)} f_1(x)\, dx = 0.$$

Thus we may essentially ignore those (x, y) for which the conditional density is not defined.

We expect that $P[Y \in C \mid X = x] = \int_C h(y \mid x)\, dy$. More generally, if $B \in \mathcal{F}$ and $X = x$, then P will assign $P[(X, Y) \in B \mid X = x]$ to the set $\{y : (x, y) \in B\}$. This we propose

$$c(y) = \int_{\{y : (x, y) \in B\}} h(y \mid x)\, dy, \qquad x \in \mathbb{R}, \quad y \in \mathbb{R},$$

as the conditional probability of B given $X = x$. To prove this, we note that

$$g(x) = \int_{-\infty}^{\infty} I_B(x, y) h(y \mid x) f_1(x)\, dy$$

being g is Borel measurable by Tonelli's theorem. Also, if $A \in \mathcal{B}(\mathbb{R})$,

$$P[(X, Y) \in (A \times \mathbb{R}) \cap B] = \int_A \int_{\{y : (x, y) \in B\}} f(x, y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_A(x) I_B(x, y) f(x, y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} I_A(x) f_1(x) \int_{\{y : (x, y) \in B\}} h(y \mid x)\, dy\, dx$$

$$= \int_A g(x)\, dx$$

$$= \int_{\mathbb{R}} g(x) I_A(x)\, dx \qquad \text{[recall } I_A(x) = 1 \text{ if } x \in A\text{]}$$

Therefore $g(x) = P[B \mid X = x]$

In this example we may look at the function $f(x,y) = f_{X|Y}(x|y)f_Y(y)$ in two ways. If (X,Y) has density f, we have a notion of conditional probability, $P(X \in A \mid Y = y)$, for any y. On the other hand, we may suppose that we specify that X has density $f_{X|Y}(\cdot|y)$ whenever $Y = y$. We could then reasonably refer to the density as $f_{X|Y}$. In other words, we specify $P(X \in A, Y \in B) = \int_B \left(\int_A f_{X|Y}(x|y) f_Y(y)\,dx \right) dy$. A unique measure P on $\mathcal{B}(\mathbb{R}^2)$ is determined, satisfying for all $A \in \mathcal{B}(\mathbb{R})$,

$$P(X \in A, Y \in B) = \int_A \left(\int_B f_{X|Y}(x|y) f_Y(y)\,dy \right) dx$$
$$= \int_B \left(\int_A f_{X|Y}(x|y) f_Y(y)\,dx \right) dy$$

Therefore (X,Y) has density $f(x,y) = f_{X|Y}(x|y) f_Y(y)$.

There are two points of view. We may regard the conditional density of X given $Y = y$ as a density constructed from the joint density of X and Y. On the other hand, we may regard the observation of X and Y as a two stage random experiment, where the distribution of Y at stage 2 depends on the value of X at stage 1. The above discussion shows that the assignment of probabilities in either way is the same in the case of two stages.

We may also define conditional densities in higher dimensions. For example, if X, Y, Z have joint density f, we define, say, the conditional density of (X,Y) given Z, by the conditional density of (X,Y) given $Z = z$ is

$$f_{(X,Y)|Z}(x,y|z) = \frac{f(x,y,z)}{f_Z(z)}$$

where $f_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y,z)\,dx\,dy$. If $B = B(\mathbb{R}^2)$, then, more generally,

$$P((X,Y) \in B \mid Z = z) = \int\int_B f_{(X,Y)|Z}(x,y|z)\,dx\,dy.$$

This is verified by showing that for $A \in \mathcal{B}(\mathbb{R})$,

$$P((X,Y) \in A \times B) = \int_B P(X \in A \mid Y = y) f_Y(y)\,dy$$

(c) Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be given, with no probability defined as yet. Take $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $X(\omega_1,\omega_2) = \omega_1$, $Y(\omega_1,\omega_2) = \omega_2$. Assume that we are given P_1, a probability measure on $(\Omega_1, \mathcal{F}_1)$, and also that we are given $P(\cdot, \cdot) : \mathcal{F}_2 \times \Omega_1$, a probability measure on $\mathcal{F}_2$ for each fixed ω_1 and a Borel measurable function of ω_1 for each fixed B. We are specifying

the distribution of X and the conditional distribution of Y, given $X = x$, by the product measure theorem. Then Σ is a unique measure P on $\mathcal{F}$ such that

$$P[Y \in A,\ F \in B] = P[A \times B] = \int_A P[x, B]\, dP_X(x).$$

It follows that $P(x, B)$ is in fact the conditional probability $P[Y \in B | X = x]$.

We now consider conditional expectation. Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. We seek to a reasonable definition of the expectation of Y given that $X = x$, written $E[Y|X = x]$. Intuitively, $E[Y|X = x]$ should reflect the long-run average value of Y in a sequence of independent trials in which we look only at those observations on which $[X = x]$ has occurred.

If X and Y are discrete and we are given that $X = x$, the conditional probability of an event involving Y is governed by the set of conditional probabilities $p(y|x) = P[Y = y | X = x]$. Thus a reasonable definition for $E[Y|X = x]$ is $\sum_y y p(y|x)$. Similarly, if (X, Y) is absolutely continuous, and $h = h(y|x)$ is the conditional density of Y given $X = x$, we expect that $E[Y|X = x]$ should be $\int_{-\infty}^{\infty} y h(y|x)\, dy$. What we need is a general framework that includes these special cases.

Let Y be a random variable (or an extended random variable) on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Our general definition of conditional probability employs a version of the theorem of loss probability

$$P[Y \in A | \mathcal{A}](\omega) = \int_\Omega P[B | \mathcal{A}](\omega)\, dP_Y(\omega), \qquad A \in \mathcal{A}^*, \qquad B \in \mathcal{F}.$$

There is a closely related theorem on "conditional expectation" which may be described informally as follows. The probability $P[Y \in B]$ is the mean value of $dP_Y(x)$, given that $X = x$. The average value of Y is what we are looking for, namely, $E[Y|X = x]$. It is reasonable to hope that the total expectation may be found by adding all the contributions.

$$E[Y] = \int_\Omega E[Y|X = x]\, dP_X(x).$$

To develop this further, we replace Y by $Y(\omega)$, where $\omega \in \Omega$. For $\omega \in A$, we suppose that $\omega(x) = \omega_{Y|X}(\omega|X = x) = E[Y|X = x]$ and $\omega(x) = 0$ if $x \notin A$. If $\omega \in A$, we expect that $E[Y|\omega \in A, X = x] = \omega(x)$. Replacing Y by $Y(\omega)$ in the above discussion of the total expectation, we obtain

$$\omega(\omega) = \int_\Omega E[Y(\omega)|X = x]\, dP_X(x)$$

or

$$\int_{\{\omega \in A\}} Y\, dP = \int_A E[Y|X = x]\, dP_X(x).$$

In fact, this requirement uniquely determines $E[Y|X = x]$.

5.3.7 *Theorem.* Let Y be an extended random variable on $(\Omega, \mathcal{F}, P)$, and $X: (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ a random object. If $\sigma(Y) \subseteq \sigma(X)$, then there is a function $g: (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$ such that for each $A \in \mathcal{F}'$

$$\int_{X^{-1}(A)} Y \, dP = \int_A g \, d(P \circ X^{-1})$$

(Actually, g is not only Borel measurable, but also has the property that it is constant on each set $X^{-1}(\{\omega'\})$, $\omega' \in \Omega'$.) We define $E(Y \mid X = \omega')$ to be $g(\omega')$ if g is essentially unique for $(P \circ X^{-1})$.

Proof. Let

$$\lambda(A) = \int_{X^{-1}(A)} Y \, dP = \int_{\{Y \geq 0\} \cap X^{-1}(A)} Y \, dP + \int_{\{Y < 0\} \cap X^{-1}(A)} Y \, dP, \qquad A \in \mathcal{F}'$$

Then λ is a countably additive set function on $\mathcal{F}'$ for which $\lambda(\emptyset) = 0$ and λ is absolutely continuous with respect to $P \circ X^{-1}$ since $P \circ X^{-1}(A) = 0$ implies $P(Y \in A) = 0$. The result follows from the Radon–Nikodym Theorem. □

Conditional expectation includes conditional probability as a special case, as we now prove.

5.3.8 *Corollary.* If Y is a random object on $(\Omega, \mathcal{F}, P)$ and $B \in \mathcal{F}$, then

$$P(B \mid X = x) = P(B \mid Y = x) \qquad [P \circ X^{-1}]$$

Proof. Let $Z = 1_B$; the definition applied to conditional expectation becomes

$$\int_{X^{-1}(A)} 1_B \, dP = \int_A P(B \mid X = x) \, d(P \circ X^{-1})(x)$$

The result now follows from $\int_{X^{-1}(A)} 1_B \, dP = P(B \cap X^{-1}(A))$. □

Let us compare the general definition with the intuitive concept we started with.

5.3.9 *Example.* (a) Let X take on only countably many values $x_1, x_2, \ldots$, assume all $P(X = x_i) > 0$, and Y be integrable. Then

$$P(B \mid X = x_i) = \frac{P(B \cap \{X = x_i\})}{P(X = x_i)} \qquad B \in \mathcal{F}$$

Thus we should expect that

$$P(B_i|X = x) = \frac{1}{P[X = x]} \int_{B_i} f_{X,Y}(x, y)\, dy.$$

Proceeding from indicators to nonnegative simple functions to nonnegative measurable functions to arbitrary integrable functions, we should like to believe that if (X, Y) exists,

$$E(g(Y)|X = x) = \frac{1}{P[X = x]} \int_{\Omega_Y} g(y) f_{X,Y}(x, y)\, dy, \qquad x = 1, 2, \ldots. \tag{1}$$

We are not proving anything here; indeed, we do not yet know, for example, that

$$P\left[\bigcup_{i=1}^{\infty} B_i \,\middle|\, X = x\right] = \sum_{i=1}^{\infty} P[B_i|X = x].$$

To establish (1), set

$$\mu(B_i) = P[X = x] \int_{B_i} g(y) f_{Y|X}(y|x)\, dy, \qquad i = 1, 2, \ldots$$

(We may assume $\Omega_Y = \{y_1, y_2, \ldots\}$, with $\mathcal{F}$ the class of all subsets of Ω_Y.) Then

$$\int_{\{x\}\times B_i} g(y)\, dP = \sum_{y\in B_i} g(y) P[X = x|Y = y] P[Y = y] = \int_{\{x\}\times B_i} g(y) f_{X,Y}(x, y)\, dy$$

$$= \sum_{y\in B_i} g(y) P[X = x|Y = y] P[Y = y] = \int_{B_i} g(y) f_{X,Y}(x, y)\, dy, \qquad B_i \in \mathcal{F},$$

as desired.

In the special case when X is discrete, (1) assumes a simple form. If X takes on the values $x_1, x_2, \ldots$, we obtain, using countable additivity of the integral,

$$E[Y|X = x_i] = \sum_j y_j \frac{P[X = x_i, Y = y_j]}{P[X = x_i]}$$

$$= \sum_j y_j P[Y = y_j | X = x_i]. \tag{2}$$

(a) Let $B \in \mathcal{F}$ and assume $P(B) > 0$. If $E(Y)$ exists, we define the conditional expectation of Y given B, as follows. Let $X = I_B$ and set $E(Y|B) = E(Y|X = 1)$. This is a special case of (b), so from Case 1 of (b)

$$E(Y|B) = \frac{1}{P(B)} \int_B Y \, dP,$$

or in other words

$$E(Y|B) = \frac{E(Y I_B)}{P(B)}. \tag{5}$$

(b) Let X and Y be random variables having a joint density f, and let $h = h(x, y)$ be the conditional density of Y given $X = x$. We assume that $E(Y)$ exists.

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y h(y|x) \, dy \tag{6}$$

To prove this, note that

$$\begin{aligned} \int_{\{X \in A\}} Y \, dP &= \int_A \int_{-\infty}^{\infty} y f(x, y) \, dy \, dx \\ &= \int_A f_1(x) \left[\int_{-\infty}^{\infty} y h(y|x) \, dy \right] dx \qquad \text{by Tonelli's theorem} \\ &= \int_{\{X \in A\}} \left[\int_{-\infty}^{\infty} y h(y|x) \, dy \right] dP(x, y) \end{aligned}$$

proving (6).

Notice also that if $g = g(x)$ is a Borel measurable function on $\mathbb{R}$ and $E[g(X)]$ exists, then

$$E(g(X) \mid X = x) = \int_{-\infty}^{\infty} g(y) h(y|x) \, dy \tag{7}$$

by the same argument as above. Similarly, if h and Y as in Exercise 11.2 (a) and $E[g(X)Y]$ exists, then

$$E(g(X)Y \mid X = x) = g(x) \int_{-\infty}^{\infty} y h(y|x) \, dy = g(x) E(Y \mid X = x). \tag{8}$$

(2) Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be given with no probability defined on either. Let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $X(\omega, \omega') = \omega$, $Y(\omega, \omega') = \omega'$. Assume that a probability measure P_1 on $\mathcal{F}_1$ is given, and also that we are given $P(\omega, B)$, $\omega \in \Omega_1$, $B \in \mathcal{F}_2$, a probability measure in B for each fixed ω, and P Borel measurable as a function of ω for each fixed B. Let P be the unique measure on $\mathcal{F}$ determined by P_1 and the $P(\omega, \cdot)$.

If $f : (\Omega_2, \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B})$ is $\mathcal{F}$-measurable and $\mathbb{E}[|f(Y)|] < \infty$, we claim that

$$\mathbb{E}[f(Y) \mid X = \omega] = \int_{\Omega_2} f(\omega', \omega) \, P(\omega, d\omega'). \tag{5.1}$$

To see this, we note, with the aid of Fubini's theorem, that

$$\begin{aligned} \int_{[X \in A]} f(Y) \, dP &= \int_{\Omega} f(Y(\omega, \omega')) \, P(d\omega, d\omega') \\ &= \int_{\Omega_1} \int_{\Omega_2} f(Y(\omega, \omega')) \, P(\omega, d\omega') \, P_1(d\omega) \\ &= \int_{\Omega_1} \int_{\Omega_2} f(\omega') \, P(\omega, d\omega') \, P_1(d\omega) \end{aligned}$$

Problems

1. Let X and Y be random variables with joint density $f(x, y)$. Indicate how to calculate the following quantities.
   (a) $\mathbb{E}[g(X) \mid Y = y]$, where $g(\cdot)$ is Borel measurable and we assume that $\mathbb{E}[|g(X)|] < \infty$;
   (b) $P(X \in A \mid Y = y)$, where $A = \{x : h(x) \in B\}$, $B \in \mathcal{B}(\mathbb{R})$;
   (c) $P(X \in A \mid Y = y)$, where $A = \{x : x + Y \in B\}$, $B \in \mathcal{B}(\mathbb{R})$.
2. Let Y have a uniform density on $[0, 1]$. If $X = k$, k a nonnegative integer, $k \le Y < k + 1$, then take $X_1, \ldots, X_n$ to be independent identically distributed observations, where each X_i has density $f_Y(x)/(k+1)$. Indicate how to compute the conditional expectation of $g(Y)$ given $X_1 = x_1, \ldots, X_n = x_n$.
3. Let X be a discrete random variable. If $X = x$, let Y have a conditional density $h(y|x)$. Show that

$$P(X = x \mid Y = y) = \frac{P(X = x) h(y|x)}{\sum_x P(X = x) h(y|x)}$$

5. Let X be an absolutely continuous random variable. If $A = \{a \le X \le b\}$, let X be discrete with $P\{X = x_i\} = P\{x_i \le X \le x_{i+1}\}$ specified. Show that there is a conditional density of X given A, namely

$$f(x|A) = \begin{cases} \dfrac{f_X(x)}{P\{A\}} & \text{if } x \in A \\ 0 & \text{if } x \notin A, \end{cases}$$

where

$$P\{A\} = P\{a \le X \le b\} = \int_a^b f_X(x)\,dx = F_X(b) - F_X(a).$$

6. (a) Let X be a discrete random variable with $P\{X = x_i\} = p_i$, $i = 1, 2, \ldots$, and let A be the event that the observed values of X are among the numbers $x_1, \ldots, x_n$. Indicate how to define $P\{X = x_i \mid A\}$.

   (b) Let X be an absolutely continuous random variable. If $A = \{a \le X \le b\}$ is the event that X is observed to lie between a and b, where each X_i has density $f(x)$, indicate how to define $P\{X = x \mid A\}$.

   (c) Let X be an absolutely continuous random variable. If $A = \{a \le X \le b\}$ is the event that X is observed to lie between a and b, where each X_i has density $f(x)$, indicate how to define $P\{X = x \mid A\}$.

   (d) Let X be an absolutely continuous random variable. If $A = \{a \le X \le b\}$ is the event that X is observed to lie between a and b, where each X_i has density $f(x)$, indicate how to define $P\{X = x \mid A\}$.

7. (a) Let X be a discrete random variable with density f_X and $A = \{Y \in B\}$, $B = \{y_1, y_2, \ldots\}$. Show that there is a conditional density for X given A, namely

$$f(x|A) = \begin{cases} \dfrac{f_X(x)}{P\{A\}} & \text{if } x \in B \\ 0 & \text{if } x \notin B. \end{cases}$$

   (b) An interpretation of the conditional density is that

$$P\{X \in B \mid A\} = \int_B f(x|A)\,dx \qquad B \subset \{x : f(x) > 0\}.$$

8. Let $\{B_1, \ldots, B_n\}$ be a mutually exclusive and exhaustive events with strictly positive probability. Establish the following version of the theorem of total expectation: If $E[X]$ exists, then

$$E[X] = \sum_{i=1}^{n} P\{B_i\} E[X \mid B_i]$$

9. Let X and Y be nonnegative random variables such that $(\Omega, \mathcal{F})$ has an induced probability measure

$$P_{XY} = \mu \times \lambda$$

where for $B \in \mathcal{B}([0,\infty)\times[0,\infty))$, $F_{X,Y}$ is the joint cdf of (X,Y) and

$$F_{X,Y}(B) = \begin{cases} 1 & \text{if } (X,Y) \in B, \\ 0 & \text{if } (X,Y) \notin B, \end{cases}$$

and

$$\mathbb{P}(B) = \int_0^\infty \int_0^\infty e^{-x} e^{-y}\, dx\, dy.$$

Thus with probability $e^{-x}e^{-y}$, (X,Y) has the probability $\frac{1}{2}$ X and Y are clearly independent, which we can verify by calculating $P(X \le x, Y \le y)$. Calculate $P(T \in B, X = x)$, $x \ge 0$, $B \in \mathcal{B}([0,\infty))$. (Hint: think first a moment.) If you observe that $X = x$, it is then certainly true that you are operating under H_0; if $X = 1$, it is absolutely certain that you are operating under H_1.

6.4 Conditional Expectation and Generalizations

It will be very convenient to regard conditional expectations as functions defined on the sample space Ω. We will first recall the main result of the previous section.

If Z is an extended random variable on $(\Omega, \mathcal{F}, P)$, where expectation exists, and $X: (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is a random variable, then $g(x) = E[Z|X = x]$ is characterized as the unique measurable function $g: (S, \mathcal{S}) \to (\overline{\mathbb{R}}, \mathcal{B})$ satisfying

$$\int_{X^{-1}(A)} Z\, dP = \int_S g(x)\, d\mu_X(x) \qquad \text{for all } A \in \mathcal{S}. \tag{1}$$

Now let $h(\omega) = g(X(\omega))$. Then $h: (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \mathcal{B})$ and $E|g(X)| < \infty$.

$$\int_{X^{-1}(A)} Z\, dP = \int_A h\, dP \qquad \text{for all } A \in \mathcal{F}. \tag{2}$$

Figure 6.4.1

Then $h(\omega)$ is the conditional expectation of Z given that X takes the value $x = X(\omega)$; consequently, it measures the average value of Z given X, but it is defined on Ω rather than S.

It will be useful to have an action on $\mathcal{F}$ by h. We claim that

$$\int_{X^{-1}(A)} h\, dP = \int_{X^{-1}(A)} Z\, dP \qquad \text{for each } A \in \mathcal{F}. \tag{3}$$

To prove (5), note that

$$\begin{aligned}\int_{f^{-1}(A)} h\,dP &= \int_\Omega g(X(\omega))\,\mathbf{1}_{f^{-1}(A)}(\omega)\,P(d\omega) \\ &= \int_{\mathbb{R}^n} g(x)\,\mathbf{1}_A(x)\,\mu_X(dx) \qquad \text{[by 4.III.3(b)]} \\ &= \int_A g(x)\,d\mu_X(x) \\ &= \int_A h(x)\,d\mu_Y(x) \qquad \text{[by (3)]}\end{aligned}$$

Since $\mathcal{F}(Y) = \{f^{-1}(A) : A \in \mathcal{B}\}$, we may express (5) as follows:

$$\int_E h\,dP = \int_A g\,d\mu_Y \qquad \text{for each } E \in \mathcal{F}(Y), \tag{6}$$

where $\mu_Y(A) = P[Y^{-1}(A)] = P[f(X) \in A]$.

The σ-field $\mathcal{F}(Y)$ will be more important for us, and we shall look at some of its properties before proceeding.

5.6.1 Definition. Let $X: (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable map. The σ-field induced by X is given by

$$\sigma(X) = X^{-1}(\mathcal{S}).$$

Thus a set in $\sigma(X)$ is of the form $\{X \in A\}$ for some $A \in \mathcal{S}$. In particular, if $X = (X_1, \ldots, X_n)$, a random vector, $\sigma(X)$ consists of all sets $\{X \in B\}$ for $B \in \mathcal{B}(\mathbb{R}^n)$.

The following result has the following properties.

5.6.2 Theorem. Let $X: (\Omega, \mathcal{F}) \to (S, \mathcal{S})$.

(a) The collection $\sigma(X)$ is the smallest subfield $\mathcal{A}$ of subsets of Ω making X measurable relative to $\mathcal{A}$ and $\mathcal{S}$.

(b) If $S = \mathbb{R}^n$, $\mathcal{S} = \mathcal{B}(\mathbb{R}^n)$, so that $X = (X_1, \ldots, X_n)$, where $X_i: (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the ith coordinate of X, then $\sigma(X)$ is the smallest σ-field $\mathcal{A}$ of subsets of Ω making each X_i measurable relative to $\mathcal{A}$ and $\mathcal{B}(\mathbb{R})$.

(c) If $X: (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$, but $(\mathbb{R}, \mathcal{B})$, then $Z = g \circ X$ for some $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$ measurable $g: (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B})$. Conversely, if $Z = g \circ X$ and $g^{-1}(\mathcal{B}) \subset \mathcal{F}$, then $Z: (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$.

Proof. (a) $\sigma(X) \subset \mathcal{F}$ since $X^{-1}(B) \in \mathcal{F}$ by definition of $\sigma(X)$; hence $\sigma(X)$ makes X measurable. If $\mathcal{G}$ is any σ-field making X measurable, $X^{-1}(B) \in \mathcal{G}$ for all $B \in \mathcal{B}$, because $\mathcal{G} \supset \sigma(X)$.

(b) Say $Y = \sum_{i=1}^{n} c_i 1_{A_i}$. Y is measurable relative to $\mathcal{G}$ and $\mathcal{F}'$ iff each A_i is measurable relative to $\mathcal{G}$ and $\mathcal{F}'$. The result follows from (a).

(c) Assume $X \ge 0$. $\mathcal{G} \mapsto \int_{\mathcal{G}} X\,dP$ is a measure on $(\Omega, \mathcal{F})$. If Z is an indicator 1_G, $G \in \sigma(Y)$, then $G = X^{-1}(A)$ for some $A \in \mathcal{B}^1$. If $f = 1_A$, then $f \circ X = 1_{X^{-1}(A)} = 1_G = Z$. If $Z = \sum_{i=1}^{n} c_i 1_{G_i}$ is a finite-valued, simple function and $G_i = X^{-1}(A_i)$, Y as above, then $Z = f \circ X$, where $f = \sum_{i=1}^{n} c_i 1_{A_i}$.

In general, let $Z_1, Z_2, \ldots$ be finite-valued simple functions such that $Z_n \uparrow Z$. Write as above $Z_n = f_n \circ X$ and, as in Lemma 5.1.1, define $f = \lim_{n \to \infty} f_n$ where the limit exists, and 0 elsewhere. Then

$$Z(\omega) = \lim_{n \to \infty} Z_n(\omega) = \lim_{n \to \infty} f_n(X(\omega)) = f(X(\omega)).$$

The measurability holds because a pointwise limit of measurable functions is measurable. $\square$

Remark. Theorem 5.1.2(c) holds equally well for a real random variable Y on a general measure space. For, as we observed at the end of Section 4.1, Y extends to arbitrary products, but since $\sigma(Y) \subset \mathcal{B}(\mathbb{R})$, see Problem 1.

Now let us return to Def. (1) at the beginning of this section:

$$\int_{G} Y\,dP = \int_{G} X\,dP \qquad \text{all } G \in \sigma(Y)$$

where $Y = \omega \to E[X \mid Y](\omega)$. Since $\mu_X(B) = P[X \in B]$ is a measure on $(\mathbb{R}, \mathcal{B}^1)$, we have $P_{\sigma(Y)}(G) = \int_G X\,dP$ by Section 4.2(vi). This fact, along with (b), characterizes Y and gives us the concept of conditional expectation given a σ-field.

5.1.3 *Theorem.* Let X be an extended random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. Assume that $E(|X|) < \infty$. Then there is a function $E[X \mid \mathcal{G}]$, $(\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B}^1)$, called the *conditional expectation of X given $\mathcal{G}$*, such that

$$\int_{G} Y\,dP = \int_{G} E[X \mid \mathcal{G}]\,dP \qquad \text{for each } G \in \mathcal{G}$$

Any two such functions are equal a.e. $[P]$. [Note that we cannot simply set $E[X \mid Y] = E[X \mid \sigma(Y)]$; $E[X \mid \mathcal{G}]$ is required to be measurable relative to $\mathcal{G}$.]

Proof. Let $\nu(G) = \int_G Y\,dP$, $G \in \mathcal{G}$. Then ν is a countably additive set function on $\mathcal{G}$. Absolutely continuous with respect to P: the result follows from the Radon Nikodym theorem. ∎

5.4.4 *Comment.* If $p(x) = \alpha(Y = x)$ and $h(x) = g(X(\cdot), x)$, then by 5.4.3, $\alpha = h(X)$, where $\mathcal{G} = \sigma(Y)$; for convenience we shall sometimes write

$$h = \alpha(\cdot \mid Y).$$

[Here is why: if $E(Y \mid X = x) = x^2$, then $E(Y \mid X) = X^2$.]

We have seen that the conditional expectation $g(x) = E(Y \mid X = x)$, $x \in \mathcal{X}$, can be interpreted to be given by the formula $h(\omega) = g(X(\omega))$, $\omega \in \Omega$ (with probability one, of course). $\mathcal{G}$ can also be a sub σ-field of $\mathcal{F}$ arising from a random object X in this way. Simply take $X : (\Omega, \mathcal{F}) \to (\Omega, \mathcal{G})$ to be the identity map $X(\omega) = \omega$, $\omega \in \Omega$. Then $X^{-1}(\mathcal{G}) = \mathcal{G}$, so if $g(x) = \alpha(Y = x)$, then $\alpha = h(X) = h(X(\cdot)) = E(Y \mid \mathcal{G})$.

Note in addition that $E(Y \mid X = x) = E(Y \mid X^{-1}(\{x\}))$ is the average value of Y given that X is known to be x. We also have a meaning for "given" $X^{-1}(\mathcal{B})$, $\mathcal{B} \subset \mathcal{S}$, where events involving X are sets of the form $\{X \in B\}$, $B \in \mathcal{B}$. And since X is the identity map, $X^{-1}(\mathcal{B}) = \mathcal{B}$. Since we don't want to distinguish between equivalent random variables, $E(Y \mid \mathcal{G})$ may be interpreted as the average value of Y given that we know, for each $G \in \mathcal{G}$, whether or not $\omega \in G$. Some examples may help to make this clear.

5.4.5 *Example.* (a) Let X be discrete, with values $x_1, x_2, \ldots$; take $\Omega = \{x_1, x_2, \ldots\}$, with $\mathcal{F}$ the class of all subsets of Ω, and assume $P(X = x_i) = p_i > 0$ for $i \geq 1$. We have seen in 5.3.3(a) that

$$g(x_i) = E(Y \mid X = x_i) = \frac{1}{P(X = x_i)} \int_{X = x_i} Y\,dP.$$

Let $h = E(Y \mid X)$, that is, $h(\omega) = g(X(\omega))$. Then h has the constant value $g(x_i)$ on the set $X = x_i$, and $\mathcal{G} = X^{-1}(\mathcal{F})$ consists of all unions of the sets $\{X = x_i\}$. Knowledge of the value of $h(\omega)$ is equivalent to knowledge, for each $G \in \mathcal{G}$, of the membership or nonmembership of ω in G.

(b) Let X and Y be random variables with a joint density f. Let $\Omega = R^2$, $\mathcal{F} = \mathcal{B}(R^2)$, $P(C) = \int_C f(x, y)\,dx\,dy$, $C \in \mathcal{B}(R^2)$, $P(\Omega) = 1$. Take $\mathcal{G} = \mathcal{F}_X$, $\mathcal{F}' = \mathcal{B}(R)$. We have seen in 5.3.5(b) that $g(x) = E(Y \mid X = x) = \int_{-\infty}^{\infty} y h_x(y)\,dy$, where h_x is the conditional density of Y given X. Let $h = g(X(\cdot))$. Then $h(\omega) = g(X(\omega))$ or $h(x, y) = g(x)$. Thus $E(Y \mid \mathcal{G})$ is constant on vertical strips, that is, $X^{-1}(\mathcal{F}')$ consists of all sets $B \times R$, $B \in \mathcal{B}(R)$. Since $(x, y) \in B$ iff $(x, y) \in B \times R$, knowledge of $X(\omega)$

processed intuitively by asking questions of the form "does $X(\omega)$ belong to B?" is equivalent to "information about membership of ω in sets of $\mathcal{A}$".

We now have a general concept of conditional probability given a model

9.4.6 **Theorem** Let $(\Omega, \mathcal{A}, P)$ be a probability space, $\mathcal{B}$ a sub-σ-field of $\mathcal{A}$, and $A \in \mathcal{A}$. There is a function $P(\cdot|\cdot) : \Omega \times \mathcal{B} \to [0,1]$ called the conditional probability of A given $\mathcal{B}$ such that

$$P(A \cap B) = \int_B P(A|\mathcal{B})(\omega)\, dP(\omega) \qquad \text{for each } B \in \mathcal{B}.$$

Any two such functions must coincide P-a.e. and it is true that

$$P(A|\mathcal{B})(\cdot) = E(I_A|\mathcal{B})(\cdot) \qquad P\text{-a.e.}$$

Proof. Let $\lambda(B) = P(A \cap B)$, $B \in \mathcal{B}$. Then λ is a countably additive finite valued set function on $\mathcal{B}$, absolutely continuous with respect to P. The existence and P-a.e. uniqueness of $P(A|\mathcal{B})(\cdot)$ follow from the Radon–Nikodým theorem. (Since λ is finite the range of $P(A|\mathcal{B})$ may be taken as $\mathbb{R}$ rather than $[0,1]$.) The connection between conditional probability and conditional expectation follows from 9.4.2 as $\lambda(B) = P(A \cap B) = \int_B I_A\, dP$.

9.4.7 **Corollary** Let $\lambda(B) = P(B|\mathcal{A}_0)$, $B \in \mathcal{A}$, a random choice. Then $\mu(B) = \int_B \lambda\, dP$ is a measure. If $h(\omega) = \mu(X^{-1}(\omega))$ then $h \circ X = \lambda$ P-a.e. by 9.4.4. If $h(\omega) = g(X(\omega))$, then $h \circ X = \lambda$ P-a.e. by 9.4.4, hence $h = P(A|\sigma(X))$ by 9.4.6. The connection is given as $P(A|X = x) = h(x)$ and $h(\omega) = g(X(\omega))$, then

$$h = P(A|\mathcal{A}_0), \qquad \text{where } \mathcal{A}_0 = \sigma(X).$$

For convenience we shall usually write

$$\lambda = P(B|\mathcal{A}).$$

Problems

1. Let $X : (\Omega, \mathcal{A}, P) \to (\Omega', \mathcal{A}')$ and $\mathcal{A}_0 \subset \mathcal{A}$, $P(X^{-1}(\cdot)) = P'(\cdot)$. We investigate conditional expectations under which $Z = f \circ X$ for some $f : (\Omega', \mathcal{A}') \to (\mathbb{R}^1, \mathcal{B}^1)$. By 9.4.2 $E(Z|\mathcal{A}_0)$ such an f can be found if $\mathcal{A}_0 = X_0^{-1}(\mathcal{A}'_0) = \sigma(X_0)$.
   (a) Assume that $\mathcal{A}'$ separates points in other words, if $\omega' \neq \omega''$, $\omega', \omega'' \in \Omega'$, $\omega' \in A$, $\omega'' \notin A$ then $A \in \mathcal{A}'$ such that the same will always hold if $\mathcal{A}'$ contains all singletons. Show that there is a function $f : \Omega \to \Omega'$ (not necessarily measurable) such that $Z = f \circ X$.
   (b) Assume that $Z = f \circ X$, where $f : \Omega' \to \mathbb{R}^1$. If $\chi(\omega) \in \mathcal{A}'$ and $\chi(\omega) = \chi(\Omega)$ consists of a single point, show that f is measurable relative to $\mathcal{A}'$ and $\mathcal{A}'$.

This by fact case (b), a reasonable $\mathcal{F}$ exists such that $Z = f \circ X$. can be found $\mathcal{F}$ separates points and $\mathcal{F}(\Omega) = \mathcal{F}$.

5.3 Properties of Conditional Expectation

The conditional expectation $E(Y|X = x)$ is a more intuitive object than the conditional expectation $E(Y|X)$, however the full force of the theory cannot be proven beyond the case in which X is a finite dimensional random vector. Thus in formal expansions in which X takes values in a Banach space, we are forced to use $E(Y|X)$.

In this section we develop the basic properties of conditional expectation in some generality for $L^1(\Omega)$ and indicate proofs only when $E(Y|X = x)$. The reader should understand conditional probabilities are obtained by replacing Y by 1_A, and see the corresponding $P(Y|X)$ are obtained by using $Y = P(A)$.

In this discussion we follow $(\Omega, \mathcal{F}, P)$ for a probability space, X a random variable on $(\Omega, \mathcal{F}, P)$, with L^1 expectation assumed to exist. $X: (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}^n)$ is a measurable object, and $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$. The phrase "$a.e.$" with no measure specified will always mean $a.e.$ P. If $Z: (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is one way that Z is $\mathcal{G}$-measurable, and $g: (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$, we say that g is $\mathcal{G}^n$-measurable.

5.3.1 Theorem. *If Y is a constant function, then*

(a) $E(Y|\mathcal{G}) = E(Y)$ *a.e.*

(a') $E(Y|X = x) = E(Y)$ *a.e.* P_X

If $Y \in L^1(P)$, then

(b) $E(E(Y|\mathcal{G})) = E(Y)$ *a.e.*

(c) $E(Y|X = x) = \int_{\mathbb{R}^n} y \, dP_{(X,Y)}(x, y)$ *a.e.* P_X.

A statement such as $E(Y|\mathcal{G}) \le E(Y'|\mathcal{G})$ is understood if Z is a version of $E(Y|\mathcal{G})$. In other words, Z satisfies the defining requirement 5.1.1, then $Z_1 \le Z$ a.e.

(a) $E(Y|\mathcal{G}) \le E(Y'|\mathcal{G})$ a.e.

(a') $E(Y|X = x) \le E(Y'|X = x)$ a.e. P_X.

Proof. (a') The function constant a is $\mathcal{G}$-measurable and

$$\int_A Y \, dP = \int_A a \, dP \qquad A \in \mathcal{G}.$$

(5.5.1) If $g(x) = a_1 + a_2 x$, then g is $\mathcal{F}$-measurable and

$$\int_{A \cap [X \in B]} g \circ X \, dP = \int_A a_1 + a_2 X \, dP.$$

(5.5.2) $\int_C g \circ X \, dP = \int_C Y \circ X \, dP$, hence

$$\int_C h(X) \mathbb{E}[Y \mid \mathcal{G}] \circ X \, dP = \int_C h(X) Y \circ X \, dP \qquad \text{for each } h \in \mathcal{B}.$$

The result follows from (5.5.1).

(5.5.3) $\int_A g \circ X \, dP = \int_{[X \in B]} Y \, dP$, hence

$$\int_A \mathbb{E}[Y; X = x] \, dP_X = \int_A \mathbb{E}[Y; X = x] \, dP_X \qquad \text{for each } A \in \mathcal{B}^1.$$

The result follows from (4.4.1).

Parts (5.5.2) and (5.5.3) follow from (5.5.1) and (5.5.4), along with the observation that $\mathbb{E}[Y \mid \mathcal{G}] \circ X$ is $\sigma[X]$-measurable.

We now prove an additivity theorem for conditional expectations.

5.5.2 Theorem. *Let $\mathcal{G} \subset \mathcal{F}$, and suppose $\mathbb{E}[Y_1] = \mathbb{E}[Y_2]$ is well defined (i.e., not of the form $\infty - \infty$). Then $\mathbb{E}[aY_1 + bY_2 \mid \mathcal{G}] = a\mathbb{E}[Y_1 \mid \mathcal{G}] + b\mathbb{E}[Y_2 \mid \mathcal{G}]$ a.e. for all $a, b \in \mathbb{R}$ and $\mathbb{E}[Y_1 \mid \mathcal{G}] + \mathbb{E}[Y_2 \mid \mathcal{G}]$ is well defined, then*

$$\mathbb{E}[Y_1 + Y_2 \mid \mathcal{G}] = \mathbb{E}[Y_1 \mid \mathcal{G}] + \mathbb{E}[Y_2 \mid \mathcal{G}] \quad \text{a.e.} \; [P].$$

Proof. (a) If $C \in \mathcal{G}$,

$$\begin{aligned} \int_C (aY_1 + bY_2) \, dP &= a \int_C Y_1 \, dP + b \int_C Y_2 \, dP \\ &= a \int_C \mathbb{E}[Y_1 \mid \mathcal{G}] \, dP + b \int_C \mathbb{E}[Y_2 \mid \mathcal{G}] \, dP \\ &\qquad \text{by the additivity theorem for integrals} \\ &= \int_C a\mathbb{E}[Y_1 \mid \mathcal{G}] \, dP + \int_C b\mathbb{E}[Y_2 \mid \mathcal{G}] \, dP \\ &\qquad \text{by definition of conditional expectation.} \end{aligned}$$

Since $\int_C a\mathbb{E}[Y_1 \mid \mathcal{G}] \, dP + \int_C b\mathbb{E}[Y_2 \mid \mathcal{G}] \, dP$ is well defined, we apply the additivity theorem for integrals:

$$\int_C (aY_1 + bY_2) \, dP = \int_C [a\mathbb{E}[Y_1 \mid \mathcal{G}] + b\mathbb{E}[Y_2 \mid \mathcal{G}]] \, dP$$

as desired.

(a') This is done as in (b), with G replaced by $\mathcal{H}$ and

$$\int_{H_i} E(Z_n \mid \mathcal{G}) \, dP \quad \text{by} \quad \int_{A} E(Z_n \mid \mathcal{H}) \, dP.$$

[illegible] the [illegible] we omit [illegible] proofs of this type with a phrase such as "Same as (c)".

The monotone convergence theorem and the fact that a nonnegative series can be integrated term by term have exact analogs for conditional expectations.

5.5.3 *Theorem.* If $Y_n \geq 0$ for all n and $Y_n \uparrow Y$ a.e., then

(a) $E(Y_n \mid \mathcal{G}) \uparrow E(Y \mid \mathcal{G})$ a.e.

(b) $E(Y_n) \uparrow E(Y)$ $\Rightarrow$ $E(Y_n \mid \mathcal{G}) \uparrow E(Y \mid \mathcal{G})$ in $L^1(\Omega)$.

Thus if $Y_n \geq 0$, then

(c) $$E\left(\sum_{n=1}^{\infty} Y_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} E(Y_n \mid \mathcal{G}) \text{ a.e.}$$

(d) $$E\left(\sum_{n=1}^{\infty} Y_n\right) = \sum_{n=1}^{\infty} E(Y_n) \Rightarrow E\left(\sum_{n=1}^{\infty} Y_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} E(Y_n \mid \mathcal{G}) \text{ a.e.}$$

In particular, if $B_1, B_2, \ldots$ are disjoint sets in $\mathcal{F}$,

(e) $$P\left(\bigcup_{n=1}^{\infty} B_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} P(B_n \mid \mathcal{G}) \text{ a.e.}$$

(f) $$P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n) \Rightarrow P\left(\bigcup_{n=1}^{\infty} B_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} P(B_n \mid \mathcal{G}) \text{ a.e.}$$

Proof. (a) $\int_A Y_n \, dP = \int_A E(Y_n \mid \mathcal{G}) \, dP \uparrow \int_A E(Y \mid \mathcal{G}) \, dP$ by 5.5.1(b), so the $E(Y_n \mid \mathcal{G})$ increase to a $\mathcal{G}$-measurable function Z by the monotone convergence theorem. $\int_A Y \, dP = \int_A Z \, dP$, because $A = E(Y \mid \mathcal{G})$ a.e.

(a') Same as (a).

(b) By 5.5.2(a), $E(\sum_{n=1}^{\infty} Y_n \mid \mathcal{G}) = \sum_{n=1}^{\infty} E(Y_n \mid \mathcal{G})$ a.e. Let $n \to \infty$ and apply part (b) of the theorem to the original sum.

(b') Same as (b).

Finally, (e) is a special case of (b), and (f) of (b').

If we take the expectation of a conditional expectation, the result is the same as if we had just taken the expectation directly. This is actually a special case of the following more general result.

5.5.4 *Theorem.* (a) $E_{\mathcal{F}}(E(Y|\mathcal{G}))=E(Y)$; hence if Y is integrable, so is $E(Y|\mathcal{G})$.

(b) $\int_F X\,dP=\int_F E(X|\mathcal{G})\,dP$ for all $F\in\mathcal{G}$.

Proof. (a) $\int_\Omega 1_\Omega\,dP=\int_\Omega E(1|\mathcal{G})\,dP$.

(b) $\int_{\Omega\setminus F} 1_F X\,dP=\int_{\Omega\setminus F} X\,dP=\int_{\Omega\setminus F} E(X|\mathcal{G})\,dP=\int_{\Omega\setminus F} 1_F E(X|\mathcal{G})\,dP$. ∎

We now prove the dominated convergence theorem for conditional expectations.

5.5.5 *Theorem.* If $|Y_n|\le Z$ for all n, where $E(Z)<\infty$, and $Y_n\to Y$ a.e., then

(a) $\lim_{n\to\infty}E(|Y_n-Y|\,|\mathcal{G})=0$ a.e.
(b) $\lim_{n\to\infty}E(Y_n|\mathcal{G})=E(Y|\mathcal{G})$ a.e.

Proof. (a) Let $Z_n=\sup_{k\ge n}|Y_k-Y|$; then $Z_n\downarrow 0$ a.e. Now $E(Z_n|\mathcal{G})$ and $E(Z|\mathcal{G})$ exist a.e. hence by 5.5.3(a), and $E(|Y_n-Y|\,|\mathcal{G})=E(|Y|\,|\mathcal{G})\le E(|Y_n-Y|\,|\mathcal{G})$ by 5.5.3(c) and 5.5.2(a), this is less than or equal to $E(Z_n|\mathcal{G})$ by 5.5.3(b). Therefore it suffices to show that $E(Z_n|\mathcal{G})\to 0$ a.e. By 5.5.3(b), $E(Z_n|\mathcal{G})\downarrow A$ a.e. for some $\mathcal{G}$-measurable function A. Since $0\le Z_n\le 2Z$, which is integrable, we have

$$0\le\int_\Omega A\,dP\le\int_\Omega E(Z_n|\mathcal{G})\,dP=\int_\Omega Z_n\,dP\to 0$$

by the unconditional convergence theorem. Thus $A=0$ a.e., as required.

(b) Same as (a). ∎

The conditional monotone convergence theorem and Fatou's lemma may be proved for conditional expectations as follows.

5.5.6 *Theorem.* Assume $Y_n\ge Z$ for all n, where $E(Z)>-\infty$.

(a) If $Y_n\uparrow Y$ a.e., then $E(Y_n|\mathcal{G})\uparrow E(Y|\mathcal{G})$ a.e.
(b) If $Y_n\downarrow Y$ a.e., then $E(Y_n|\mathcal{G})\downarrow E(Y|\mathcal{G})$ a.e.
(c) $\liminf_{n\to\infty}E(Y_n|\mathcal{G})\ge E(\liminf_{n\to\infty}Y_n|\mathcal{G})$ a.e.
(d) $\limsup_{n\to\infty}E(Y_n|\mathcal{G})\le E(\limsup_{n\to\infty}Y_n|\mathcal{G})$ a.e.

Now assume $Y_n\le Z$ for all n, where $E(Z)<+\infty$.

(a) If $X_n \downarrow X$ a.s., then $E(X_n \mid \mathcal{F}) \downarrow E(X \mid \mathcal{F})$ a.s.
(b) If $X_n \to X$ a.s., then $E(\sup_{m \geq n} |X_m - X| \mid \mathcal{F}) \downarrow 0$ a.s. $[P_\nu]$.
(c) $\limsup_{n \to \infty} E(X_n \mid \mathcal{F}) \leq E(\limsup_{n \to \infty} X_n \mid \mathcal{F})$ a.s.
(d) $\liminf_{n \to \infty} E(X_n \mid \mathcal{F}) \geq E(\liminf_{n \to \infty} X_n \mid \mathcal{F})$ a.s. $[P_\nu]$.

Proof. (a) If $A \in \mathcal{F}$, then $\int_A X_n \, dP = \int_A E(X_n \mid \mathcal{F}) \, dP$, and $E(X_n \mid \mathcal{F})$ is nonincreasing and bounded below by $E(X \mid \mathcal{F})$. By the ordinary monotone convergence theorem, $\int_A X_n \, dP \to \int_A Y \, dP$, and therefore $Y = E(X \mid \mathcal{F})$ a.s.

(b) Let $Z_n = \sup_{m \geq n} |X_m - X|$; we then have $Z_n \downarrow 0$ and $|E(X_n \mid \mathcal{F}) - E(X \mid \mathcal{F})| \leq E(Z_n \mid \mathcal{F})$ a.s. But $0 \leq Z_n \leq 2$, so

$$\begin{aligned} E(\inf_n E(Z_n \mid \mathcal{F})) &= \lim_{n \to \infty} E(E(Z_n \mid \mathcal{F})) = \lim_{n \to \infty} E(Z_n) \\ &= \liminf_n E(Z_n \mid \mathcal{F}) \quad \text{by (a)} \\ &= 0 \quad \text{a.s.} \end{aligned}$$

(c) This follows from (a) upon replacing X_n by $\sup_{m \geq n} X_m$.

(d) This follows from (c) upon replacing X_n by $-X_n$.

Lemma (a) $E(\liminf_n P_\nu(\cdot \mid \mathcal{F}) \mid \mathcal{F}) = \liminf_n P_\nu(\cdot \mid \mathcal{F})$ a.s. $[P_\nu]$.

$$\begin{aligned} E(\liminf_n P_\nu(\cdot \mid \mathcal{F}) \mid \mathcal{F}) &\leq \liminf_n E(P_\nu(\cdot \mid \mathcal{F}) \mid \mathcal{F}) \\ &= \liminf_n P_\nu(\cdot \mid \mathcal{F}) \quad \text{by (a)} \\ &= \sup_{n \geq 0} \inf_{m \geq n} P_\nu(\cdot \mid \mathcal{F}). \end{aligned}$$

The proofs of (b) and (c) are the same as the proofs of (a) and (d). ∎

Now for our next assumption: $P(\cdot \mid \mathcal{F})$ and $E(\cdot \mid \mathcal{F})$ are under various hypotheses on P. Now we impose conditions on X and $\mathcal{A}$.

5.5.7 Theorem (a) $\mathcal{A}(P_\nu \mid \mathcal{F}) = \mathcal{A}(P \mid \mathcal{F})$ a.s.

(a) If Z is a constant k a.s., then $E(T \mid \mathcal{F}) = k = E(T)$ a.s. $[P_\nu]$.
(b) $E(T \mid \mathcal{F}) = T$ a.s.
(c) If $\mathcal{F}(\Omega, \mathcal{F}) \to (\mathcal{Z}, \mathcal{Z})$ is the measurable map $\omega \to E(T \mid \mathcal{F})(\omega)$, then $P(T^{-1}(Z) \mid \mathcal{F}) = 1_{Z \circ T}$ a.s. $[P]$.

Lemma (a) $\int_A X \, dP = \int_A E(T \mid \mathcal{F}) \, dP$ if $A = \omega$ or Ω.
(b) If $\nu \ll P$,

$$\int_{T^{-1}(\cdot)} X \, dP = \int_{(\cdot)} T \, dP = E(T) = \int_{(\cdot)} E(T) \, dP_\nu.$$

If $\nu \ll P$,

$$\int_{T^{-1}(\cdot)} T \, dP = \int_{(\cdot)} T \, dP = 0 = \int_{(\cdot)} E_\nu(T) \, dP_\nu.$$

(b) $\int_A Y\,dP = \int_A X\,dP$ for all $A \in \mathcal{B}$, and Y is $\mathcal{B}$-measurable.
(c) $\int_{\{Y>c\}} Y\,dP = \int_{\{Y>c\}} X\,dP$ and Y is $\mathcal{B}_{\{Y>c\}}$-measurable. □

The following result is unexpectedly the most important.

5.5.8 *Theorem.* If $(\Omega, \mathcal{A}, P)$ is a measure space, $\mathcal{B} \subset \mathcal{A}$ is a sub-σ-algebra, and $B \in \mathcal{B}$ is an atom of $\mathcal{B}$, that is, $B \neq \emptyset$, $P(B) > 0$, and $A \in \mathcal{B}$, $A \subset B$ then $P(A) = 0$ or $P(B - A) = 0$, then f is an a.e. constant on B.

Proof. If $c \in \mathbb{R}$ and $\{x : f(x) < c\} \cap B = \emptyset$, then $f(x) \geq c$ a.e. on B. Let $c = \sup\{r \in \mathbb{R} : P(\{x \in B : f(x) < r\}) = 0\}$. Then

$$P(\{x \in B : f(x) < c\}) = P\left(\bigcup_{n=1}^{\infty} \{x \in B : f(x) < c_n\}\right) = 0.$$

Hence $P(\{x \in B : f(x) < c\}) > 0$. Hence $P(\{x \in B : f(x) \geq c\}) = 0$ since B is an atom. Thus

$$P(\{x \in B : f(x) > c\}) = P\left(\bigcup_{n=1}^{\infty} \{x \in B : f(x) > c_n\}\right) = 0$$

so $f = c$ a.e. on B. □

We now show that conditional expectation is an "averaging" or "smoothing" operation. If $\mathcal{B}$ is an atom of $\mathcal{A}$ $E(|X|) < \infty$ and $B \in \mathcal{B}$, $P(B) > 0$, then $E(X|\mathcal{B})$ is the average value of X on B.

5.5.9 *Theorem.* (a) Let B be an atom of $\mathcal{B}$ relative to P. Then

$$E(X|\mathcal{B}) = \frac{1}{P(B)} \int_B X\,dP = \frac{E(X I_B)}{P(B)} \quad \text{a.e. on } B.$$

As a special case, let $B_1, B_2, \ldots$ be the atoms of $\mathcal{B}$ whose union is Ω, with $P(B_n) > 0$ for all n. Let $\mathcal{B}$ be the σ-field over the B_n, so that Y is the collection of all unions of sets from the B_n. Then

$$E(X|\mathcal{B}) = \sum_{P(B_n)>0} \frac{1}{P(B_n)} \int_{B_n} X\,dP \quad \text{a.e. on } B_n, \quad n = 1, 2, \ldots.$$

(b) If $B = \{X = a\}$ and $P(B) > 0$, then

$$E(X|X = a) = \frac{1}{P(B)} \int_B X\,dP.$$

PROOF (a) By (5.5.8), $\int_A Y dP = \int_A X dP$ for all $A \in \mathcal{G}$. Since $\int_A Y dP = \int_A E(Y|\mathcal{G}) dP$ for all $A \in \mathcal{G}$ it is unique. (b) If $X = 1_B$, we take $E(1_B|\mathcal{G}) = P(B|\mathcal{G})$, as desired.

(c) We first show that $\mathcal{B}$ is an algebra of $\mathcal{B} = \sigma(\mathcal{C})$. For measurable $A_1, \ldots, A_n$ and $X = \prod_{i=1}^n 1_{A_i}$ if $\tau_k = k$, then $X^{-1}\{\eta_k = 1\} = X^{-1}\{\eta_k = 1\}$ because $X^{-1}\{1\} = \bigcap_{i=1}^n A_i \in \mathcal{B}$. Also $X^{-1}\{0\} \cap X^{-1}\{\eta_k = 1\} = \emptyset$, so $X^{-1}\{\eta_k = 0\} = X^{-1}\{1\}^c$. Thus $X^{-1}\{\eta_k = 0\} = X^{-1}\{1\}^c - X^{-1}\{\eta_k = 1\} \in \mathcal{B}$, so $X^{-1}\{\eta_k = 0\} \in \mathcal{B}$ because $X^{-1}\{1\} \in \mathcal{B}$.

Now let $g(x) = E(Y|X = x)$, $h(\omega) = g(X(\omega)) = E(Y|X)(\omega)$. By (a), $\int_A h dP = \int_A Y dP$ for $A = X^{-1}(B)$, $B \in \mathcal{B}$. This is easily seen to hold for all $A \in \mathcal{B}$. Thus h is a version of $E(Y|\mathcal{B})$. If τ is any other function such that $\int_A \tau dP = \int_A Y dP$ for all $A \in \mathcal{B}$, then $g(X(\omega)) = g(\eta_k) = \tau$ is constant on $\{\omega : \tau_k = k\}$ $\square$

We now consider successive conditioning relative to two σ-fields, one of which is nested within the other. That is, a version of $E(Y|\mathcal{G})$ is given. The result is true no matter what conditioning operation is applied first, the result is the same as if only the conditioning with respect to the smaller σ-field is done. This is intuitively reasonable. For example, to find the average value of a real-valued function f defined on $[0, 3]$, we may compute the average of f on $[0, 1]$ and on $[1, 3]$, then the average of f on $[0, 1]$ and $[1, 3]$. The average of f on $[0, 3]$ is $\frac{1}{3}\int_0^3 f(x)\,dx$, which is then $\frac{1}{3}a_1 + \frac{2}{3}a_2$.

(5.5.10) THEOREM (a) If $\mathcal{G}_1 \subset \mathcal{G}_2$ then $E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1)$ a.e.

(b) If $\mathcal{F}_1 \subset \mathcal{F}_2$ and $E|X| < \infty$, then $E(E(X|\mathcal{F}_2)|\mathcal{F}_1) = E(X|\mathcal{F}_1)$ a.e.

(c) If $\mathcal{F}_1 \subset \mathcal{F}_2$ and $E|X| < \infty$, then $E(E(X|\mathcal{F}_2)|\mathcal{F}_1) = E(X|\mathcal{F}_1)$ a.e.

PROOF (a) If $G \in \mathcal{G}_1$, then $\int_G E(E(X|\mathcal{G}_2)|\mathcal{G}_1)\,dP = \int_G E(X|\mathcal{G}_1)\,dP = \int_G E(X|\mathcal{G}_2)\,dP$ since $G \in \mathcal{G}_2$. Thus $E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1)$ a.e. Conversely, if $G \in \mathcal{G}_1$, then $\int_G E(E(X|\mathcal{G}_2)|\mathcal{G}_1)\,dP = \int_G E(X|\mathcal{G}_1)\,dP = \int_G E(X|\mathcal{G}_2)\,dP$ since $G \in \mathcal{G}_2$. Thus $E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(E(X|\mathcal{G}_1)|\mathcal{G}_2)$ a.e.

(b) Let $\mathcal{G}_1 = \{\emptyset, \Omega, A, A^c\}$, $A \in \mathcal{F}_1$, $\mathcal{G}_2 = \mathcal{F}_1$, and apply (a).

(c) $E(X|\mathcal{F}_1)$ is $\mathcal{F}_1$-measurable, hence $\mathcal{F}_2$-measurable, and

$$\int_A E(X|\mathcal{F}_1)\,dP = \int_A E(X|\mathcal{F}_2)\,dP, \qquad A \in \mathcal{F}_1$$

(5.5.7). Take $\mathcal{F}_2$ and $\mathcal{F}_1$ as in (a) and apply (5.5.7) $\square$

If we take the conditional expectation of a product of two random variables, under certain conditions one of the factors can be factored out, as follows.

(5.5.11) THEOREM (a) If Z is $\mathcal{F}$-measurable and both Y and YZ are integrable, then $E(YZ|\mathcal{F}) = ZE(Y|\mathcal{F})$ a.e. In particular, $E(Z|\mathcal{F}) = Z$ a.e.

(c) If $f : (\Omega^2, \mathcal{F}^2) \to (\mathbb{R}^k, \mathcal{R}^k)$ and both X and Y or $f(X)$ are integrable then $E(f(X)Y|\mathcal{G}) = f(X)E(Y|\mathcal{G})$ a.s. [P_X]. In particular

$$E(f(X)|X = x) = f(x) \quad \text{a.e. } [P_X].$$

Proof. (a) If Z is an indicator I_B, $B \in \mathcal{F}$, and $C \in \mathcal{G}$ we have

$$\int_C Z\,dP = \int_{C\times S} I_B\,dP = \int_{C\times\Omega} E(Z|\mathcal{G})\,dP = \int_C I_B E(Y|\mathcal{G})\,dP$$
$$\int_C Y_1(Y, \mathcal{G})\,dP,$$

where $Y_1(Y, \mathcal{G})$ is $\mathcal{G}$-measurable. Thus the result holds for indicators.

Now let Z be simple, say

$$Z = \sum_{j=1}^{n} z_j I_{B_j} \qquad \text{with} \qquad B_j \in \mathcal{F}.$$

By 5.5.2(a),

$$E(|Z|\mathcal{G}) = \sum_{j=1}^{n} |z_j| E(|I_{B_j}||\mathcal{G}) = \sum_{j=1}^{n} |z_j| I_{B_j} E(|I_\Omega|\mathcal{G}) = |Z| E(I_\Omega|\mathcal{G}).$$

If Z is an arbitrary $\mathcal{F}$-measurable function, let $Z_1, Z_2, \ldots$ be simple functions $\mathcal{F}$-measurable with $|Z_n| \le |Z|$ and $Z_n \to Z$. Now $E(|Z_n|\mathcal{G}) \le |Z| E(I_\Omega|\mathcal{G})$ by what we have just proved, and $E(|Z_n|\mathcal{G}) \to E(|Z|\mathcal{G})$ by 5.5.2(c). (The integrability of $|Z|$ is used here.) Since Z is integrable so is $E(|Z|\mathcal{G})$; hence $E(|Z|\mathcal{G})$ is finite a.s., and consequently $|Z_n| E(|I_\Omega|\mathcal{G}) \to |Z| E(|I_\Omega|\mathcal{G})$ a.s. Note that, for example, if $Z = 0$ but $E(|Z||\mathcal{G}) = \infty$, then finiteness of $E(|Z|\mathcal{G})$ is important. Therefore, $E(|Z|\mathcal{G}) = |Z| E(|I_\Omega|\mathcal{G})$.

(b) Let $Z = I_B$, $B \in \mathcal{F}^1 \otimes \mathcal{F}^2$. Then

$$\int_{A\times\Omega} Y(X, \cdot)\,dP = \int_{A\times\Omega} I_B(X, \cdot)\,dP = \int_{A\times\Omega} Y\,dP$$
$$= \int_{A\times\Omega} E(Y|X = x, \cdot)\,dP^{X,Y}(x, \cdot) = \int_A g(x)\,dP^X(x) = \int_{A\times\Omega} g(X)\,dP$$

Thus the result holds when f is an indicator. Passage to simple functions and then to arbitrary measurable functions is carried out just as in (a). ∎

5.5.3 (Comments on (a)) Theorem 5.5.1(a) holds under the weaker assumption that $P(Y \text{ is real}) = P(Y \ge 0) = 1$ and $E(Y) = \infty$ [or $P(Y \text{ is real}) = 1$ and $E(Y)$ exists]; see Problem 4.

(b) *Intuitively* Is it known that $X = x$, then $g \circ X$ may be treated as a constant. That is, in the $E(Y; f \circ X)|X = x)$ should be $f(x)E(Y|X = x)$ as in 3.3.1(b). A similar result may be given in 5.3.1(b) as follows. We can express $E(g \circ X|\mathcal{C})$ as $h(X)$ where $h(x) = \int g(x)\,dF_{Y|X}(y|x)$ for an appropriate random object λ (see 5.7.A). Since Z is $\sigma(X)$ measurable, $Z = f \circ X$ for some $f : (\mathbb{R}, \mathcal{B}^1) \to (\mathbb{R}, \mathcal{B}^1)$ by 5.A.2(c). Thus if $X = x$ is given, Z may be treated as a constant and factored out.

Problems

1 (a) If X is a random object, Y a random variable such that $\sigma(Y)$ and $\sigma(X)$ are independent, show that $E(Y|X) = E(Y)$ a.e. $[P_X]$ and $E(Y|X = x) = E(Y)$ a.e. $[P_Y]$.

(b) Give an example of integrable random variables X and Y such that $E(Y|X) = E(Y)$, but X and Y are not independent.

2 If X is an $\mathbb{R}^n$-valued random variable and X and Z are random objects, show that $\sigma(X, Z)$ and Z are independent if and only if $P_{(X,Z)} = P_X \times P_Z$.

3 This problem illustrates how to obtain a theorem about $E(Y|X = x)$ from a corresponding theorem about $E(Y|X)$ without writing a separate proof or relying "similarity." As a typical example, suppose we have proved that if $Y_1 \le Y_2$ a.e., then $E(Y_1|X) \le E(Y_2|X)$ a.e. Show that

$$P(\{x : E(Y_1|X = x) > E(Y_2|X = x)\}) = P(\{\omega : E(Y_1|X)(\omega) > E(Y_2|X)(\omega)\}) = 0.$$

Conclude that if $Y_1 \le Y_2$ a.e., then $E(Y_1|X = x) \le E(Y_2|X = x)$ a.e. $[P_X]$.

4 Prove Theorem 5.3.1(a) in the case where $A(P)$ and $\sigma(P)$ need not be independent.

5 Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. If $X \in L^2(\Omega, \mathcal{F}, P)$, then $E(X|\mathcal{G})$ is also in $L^2(\Omega, \mathcal{F}, P)$, by 5.3.1(a). Thus $A(X) = E(X|\mathcal{G})$ defines a linear operator on L^2, and it may be considered in the Hilbert space $L^2(\Omega, \mathcal{F}, P)$.

(a) Show that $\|A\| = 1$.

(b) Define $\langle X, Y \rangle = \int XY\,dP$, $X, Y \in L^2$. Show that A has the "self-adjointness" property $\langle AX, Y \rangle = \langle X, AY \rangle$. [Note that $AY \in L^2$ and if $X \in L^2$, then $AX \in L^2$, so $\langle X, AY \rangle$ is well-defined.]

5.4 Regular Conditional Probabilities

We have seen that if $B_1, B_2, \ldots$ are disjoint sets in $\mathcal{F}$, $P(B) > 0$, then

$$P\left(\bigcup_{n=1}^{\infty} B_n | B\right) = \sum_{n=1}^{\infty} P(B_n|B) \qquad (*)$$

This does not imply that we will be able to find a set $B(B)\in\mathcal{F}$ for $B\in\mathcal{M}$ such that $B(B)$ is measurable in $\mathcal{B}$ for all ω and all $B\in\mathcal{B}$. To satisfy the problem, suppose we for each $B\in\mathcal{B}$ have chosen a particular version of the conditional probability $P(B|\mathcal{F})(\omega)$, $\omega\in\Omega$; we now have a number $P(B|\mathcal{F})(\omega)$ for each $B\in\mathcal{B}$ and $\omega\in\Omega$. The difficulty is that for a fixed ω, $P(\cdot|\mathcal{F})(\omega)$ need not be a countably additive set function. Suppose, for example, that $\{B_1, B_2, \ldots\}$ are disjoint sets in $\mathcal{B}$. Then

$$P\left[\bigcup_{n=1}^{\infty} B_n \,\middle|\, \mathcal{F}\right](\omega) = \sum_{n=1}^{\infty} P(B_n|\mathcal{F})(\omega)$$

except for ω belonging to a set $N(B_1, B_2, \ldots)$ of probability 0. Thus the set N of ω's for which countable additivity fails is

$$N = \bigcup \{N(B_1, B_2, \ldots) : B_1, B_2, \ldots \text{ disjoint sets in } \mathcal{B}\}.$$

In general, N is an uncountable union of sets of probability 0, and therefore N may not have probability 0 even when P is. Thus we have no guarantee that we can specify the $P(B|\mathcal{F})$ to be countably additive in B. Similarly, there is no guarantee that $P(B|\mathcal{F})=1$ will be countably additive in B for B outside all P_θ [

We are going to establish conditions under which the countable additivity requirement can be met.

6.6.1 Definition. Let T be a random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. The function $K = K(\omega, B)$, $\omega\in\Omega$, $B\in\mathcal{B}$, is called a *regular conditional distribution function for T given $\mathcal{G}$* if the following two conditions are satisfied.

1. For each ω, $K(\omega, \cdot)$ is a proper distribution function on $\mathcal{B}$, that is, nondecreasing and right continuous, with $K(\omega, -\infty) = \lim_{x\to-\infty} K(\omega, x) = 0$, $K(\omega, \infty) = 1$.
2. For each x, $K(\cdot, (-\infty, x]) = P[T \le x \,|\, \mathcal{G}](\omega)$ for almost every ω.

The key point in the development is the result that a regular conditional distribution function for a given random variable always exists.

6.6.2 Theorem. If T is a random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$, there always exists a regular conditional distribution function for T given $\mathcal{G}$.

Proof. Select a version of $P[T \le x \,|\, \mathcal{G}](\omega)$ for each fixed rational x, and let $\{r_1, r_2, \ldots\}$ be an enumeration of the rationals. Let

$$\alpha_{ij} = \inf\{P[T \le r_j \,|\, \mathcal{G}](\omega) : \omega \in A_i\}, \qquad A_i = \bigcap_{j=1}^{\infty} \{\omega : P[T \le r_j \,|\, \mathcal{G}](\omega) \ge \alpha_{ij}\}.$$

Then $P(A)=0$ since $r_n \le y$ implies $F(x) \le F(y) \le F(r_n)$, i.e., by (5.1.1).

Now let

$$N = \left\{\omega : \lim_{r \in \mathbb{Q}} F_{X,Y}(r,y) \text{ does not exist}\right\}, \qquad B = \bigcup_{n=1}^{\infty} B_n$$

Since

$$\lim_{x \to -\infty} F(x,y) = 0 \qquad \text{as} \qquad x \to -\infty$$

(5.1.3) yields

$$\lim_{r \in \mathbb{Q},\, r \to -\infty} F(r,y) = 0 \qquad \text{for} \qquad \omega \in B^c$$

hence $P(B)=0$.

If $\Gamma = \{\omega : \lim_{r \in \mathbb{Q},\, r \to \infty} F_X(r) = 1\}$, then $P(\Gamma)=0$ since $\{X \le r\} \uparrow \Omega$, so that $P(\{X \le r\}) \to 1$ as $r \to \infty$. Similarly, $\Delta = \{\omega : \lim_{r \in \mathbb{Q},\, r \to -\infty} F_X(r) \ne 0\}$ is negligible, i.e., $P(\Delta)=0$.

Define

$$F(x,y) = \begin{cases} \lim_{r \in \mathbb{Q},\, r \downarrow x} F(r,y) & \text{if } \omega \notin N \cup B \cup \Gamma \cup \Delta \\ \text{any proper distribution function } G(x,y) & \text{if } \omega \in N \cup B \cup \Gamma \cup \Delta \end{cases}$$

Then F is well defined, for if $\omega \notin N$, then $F(x,y)$ is nondecreasing in x, so that $\lim_{r \in \mathbb{Q},\, r \downarrow x} F(r,y)$ exists. Next we show that if $\omega \notin N \cup B$, then $\lim_{x \to -\infty} F(x,y) = 0$. Now if y is rational, say $y = y_n$, then $\lim_{x \to -\infty} F(x,y) = \lim_{r \to -\infty} F(r,y_n) = 0$. Similarly if $\omega \notin N \cup B \cup \Gamma \cup \Delta$, then $\lim_{x \to \infty} F(x,y) = 1$. Finally, $\lim_{y \to -\infty} F(x,y) = 0$.

We show that F is a regular conditional distribution function for P given $\mathcal{X}$. First, if $\omega \notin B \cup \Gamma \cup \Delta$, then $F(\omega;x,y)$ is clearly nondecreasing. If $x_1 < x_2$, then $F(\omega;x_1,y) = \lim_{r \in \mathbb{Q},\, r \downarrow x_1} P(\{X \le r, Y \le y\} \mid \mathcal{X})(\omega) \le \lim_{r \in \mathbb{Q},\, r \downarrow x_2} P(\{X \le r, Y \le y\} \mid \mathcal{X})(\omega) = F(\omega;x_2,y)$ since $P(\{X \le r, Y \le y\} \mid \mathcal{X})(\omega)$ is nondecreasing in r. Thus $F(\omega;x,y)$ is right continuous. If $x_n \downarrow x$, then $F(\omega;x_n,y) \downarrow F(\omega;x,y) = \lim_{r \in \mathbb{Q},\, r \downarrow x} F(\omega;r,y)$ as $n \to \infty$ because $F(\omega;r,y)$ is right continuous in x. Also $F(\omega;x,y) \to 1$ as $x \to \infty$, similarly $F(\omega;x,y) \to 0$ as $x \to -\infty$. Thus the first requirement is satisfied.

Now $P(\{X \le r\} \mid \mathcal{X})(\omega) = F_X(r) = F(\omega;r,\infty)$ by comparison of Γ. As $r \downarrow x$, $F(\omega;r,\infty) \to F(\omega;x,\infty)$ so that all $x \in \mathbb{R}$ are continuity, and

$$P\{X \le x\} \mid \mathcal{X})(\omega) = F(\omega;x,\infty) \quad \text{a.s.}$$

by (5.3.5). Thus $P(\{X \le x\} \mid \mathcal{X})(\omega) = F(\omega;x,y)$ for almost every ω for each x, which establishes the second requirement. ∎

We are going to show that if Y is a random variable, $P(Y \in B \mid \mathcal{X})(\omega)$ can be chosen to be a measure on $\mathcal{B}$ for almost every ω. This will follow from (5.3.2) and

that such that a distribution function determines a unique Lebesgue–Stieltjes measure.

5.6.3 *Definition.* Let $Y : (\Omega, \mathcal{F}, P) \to (\Omega', \mathcal{F}')$ be a random object and $\mathcal{H}$ a sub-σ-field of $\mathcal{F}$. The function $Q : \Omega \times \mathcal{F}' \to [0,1]$ is called a *version of conditional probability for Y given $\mathcal{H}$* iff

(1) $Q(\omega, \cdot)$ is a probability measure on $\mathcal{F}'$ for each fixed $\omega \in \Omega$, and

(2) for each fixed $B \in \mathcal{F}'$, $Q(\omega, B) = P[Y^{-1}B \mid \mathcal{H}](\omega)$ a.e.

If Y is a random variable, so that $\Omega' = \mathbb{R}$, $\mathcal{F}' = \mathcal{B}(\mathbb{R})$, a regular conditional probability for Y given $\mathcal{H}$ always exists.

5.6.4 *Theorem.* Let Y be a random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{H}$ a sub-σ-field of $\mathcal{F}$. There exists a regular conditional probability for Y given $\mathcal{H}$.

Proof. Let F be a regular conditional distribution function for Y given $\mathcal{H}$. Define

$$Q(\omega, B) = \int_{(\omega,\infty)} dF(\omega, y).$$

Then for each ω, $Q(\omega, \cdot)$ is the Lebesgue–Stieltjes measure corresponding to $F(\omega, \cdot)$; hence Q is a probability measure in B if ω is fixed.

Now let $\mathcal{C} = \{B \in \mathcal{B}(\mathbb{R}) : Q(\omega, B) = P[Y^{-1}B \mid \mathcal{H}](\omega)$ a.e.$\}$. Then $\mathcal{C}$ contains all intervals of the form $(\alpha, \beta]$ since $F(\omega, y) = P[Y \le y \mid \mathcal{H}](\omega)$ a.e. If $A_n \in \mathcal{C}$, $A_n \uparrow A$, then $A = \bigcup_n A_n \in \mathcal{C}$, and if $B \in \mathcal{C}$ contains all intervals $(a, b]$ hence all Borel sets, since all Borel sets are disjoint unions or right-semiclosed intervals. By the monotone class theorem, $\mathcal{C} = \mathcal{B}(\mathbb{R})$. Thus Q is a regular conditional probability for Y given $\mathcal{H}$.

We now extend this result to objects Y more general than random variables.

5.6.5 *Theorem.* Let $Y : (\Omega, \mathcal{F}, P) \to (\Omega', \mathcal{F}')$ be a random object, and $\mathcal{H}$ a sub-σ-field of $\mathcal{F}$. Suppose there is a map $\Phi : (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that Φ is one-to-one, $\mathcal{F}' = \Phi^{-1}(\mathcal{B}(\mathbb{R}))$ (i.e. all sets of $\mathcal{F}'$ are measurable), and $\Phi^{-1} : (\Phi(\Omega'), \mathcal{B}(\Phi(\Omega'))) \to (\Omega', \mathcal{F}')$ is measurable. Then there is a regular conditional probability for Y given $\mathcal{H}$.

Proof. Let $Q_0 = Q_0(\omega, B)$, $B \in \mathcal{B}(\mathbb{R})$, $\omega \in \Omega$, be a regular conditional probability for the random variable $\Phi(Y)$ given $\mathcal{H}$. Define $Q(\omega, A) = Q_0(\omega, \Phi(A))$, $A \in \mathcal{F}'$; since Φ^{-1} is measurable, $\Phi(A) = \{\Phi(\omega) : \omega \in A\} \in \mathcal{B}(\mathbb{R})$, and Q is well-defined. Since Q_0 is a probability measure in its second argument, so is Q if ω is fixed. Also

$$Q(\omega, A) = P[\omega : Y(\omega) \in A \mid \mathcal{H}](\omega) = P[Y \in A \mid \mathcal{H}](\omega) \quad \text{a.e.} \quad \square$$

A map Φ of the type described in Ex. 6.3.1 is called a Skorokhod embedding of μ into $\mathcal{L}$.

Problems

1. Let $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{R}^{[0,\infty)}, \mathcal{B}^{[0,\infty)}, \mathbb{P}^x)$ and $X = (X_t, \mathcal{F}_t)_{t \geq 0}$ be the canonical process on $(\Omega, \mathcal{F}, \mathbb{P})$, and assume that $\mathbb{P}^x(B) = \int_B p(t, x, y)\, dy$, $B \in \mathcal{B}$, where p is a probability density function on $\mathbb{R}^n \times \mathbb{R}^n \times (0, \infty)$ and $\mathbb{R}^n \times \mathbb{R}^n \times (0, \infty) \ni (x, y, t) \to p(t, x, y)$ is Borel measurable.

   (a) Show that $\mathbb{P}^x(X_{t+s} \in B \mid \mathcal{F}_t) = \mathbb{P}^{X_t}(X_s \in B)$ $\mathbb{P}^x$-a.s. [Hint]
   In particular, if $\Gamma \in \mathcal{B}^{[0,\infty)}$, then

   (b) $\mathbb{P}^x((X_{t+s})_{s \geq 0} \in \Gamma \mid \mathcal{F}_t) = \mathbb{P}^{X_t}((X_s)_{s \geq 0} \in \Gamma)$ $\mathbb{P}^x$-a.s. [Hint]
   Conclude that

   (c) $\mathbb{P}^x((X_{t+s})_{s \geq 0} \in \Gamma \mid \mathcal{F}_t) = \int_{\mathbb{R}^n} \mathbb{P}^y((X_s)_{s \geq 0} \in \Gamma)\, \mathbb{P}^x(X_t \in dy)$.

2. Let $\xi_1, \ldots, \xi_n$ be independent random variables, each uniformly distributed on $[0, 1]$, and let $\xi_{1,n}$ be the kth largest of $\xi_1, \ldots, \xi_n$, $1 \leq k \leq n$. Show that the joint density of $(\xi_{1,n}, \ldots, \xi_{n,n})$ is uniformly bounded on $\mathbb{R}^n$ and conclude that

$$\mathbb{P}(\xi_{1,n} \in B \mid \xi_{2,n} = x_2, \ldots, \xi_{n,n} = x_n) \leq \frac{1}{x_2} \int_B g(x_2)(x_2 - x_1)\, dx_1$$

   where $g(x_2) = \sup_{x_3, \ldots, x_n} g(x_2, \ldots, x_n)$ and g is the density function. [There is another way of looking at Example 6.3.2.]

3. The following result is preparatory to the next problem. It is adapted from Halmos (1950).

   (a) Let U be a Lebesgue measurable subset of $\mathbb{R}$ such that $0 < \mu(U) < \infty$, where μ is Lebesgue measure. If $0 < \alpha < 1$, show that there is an open interval I such that $\mu(U \cap I) > \alpha \mu(I)$.

   (b) With U as in part (a), let $D(U) = \{x : \text{for some } \varepsilon > 0,\ \mu(U \cap (x - \varepsilon, x + \varepsilon)) = 2\varepsilon\}$. Show that $D(U)$ includes a non-trivial interval $(0, \delta)$. [This holds also if $\mu(U) = \infty$, since $U = \bigcup_{n=1}^{\infty} (U \cap [n-1, n])$ and $\mu(U \cap [n-1, n]) > 0$ for some n.]

   (c) Let q be an irrational number, and let $A = \{m + nq : m, n \text{ integers}\}$. Show that A is dense in $\mathbb{R}$. Equivalently, the set of numbers of the form $m + nq$, m an integer, is dense in $[0, 1]$. [Assume not. Then $[0, 1] \setminus A$ is uncountable.]

with irrational ratio and let $\alpha \in [0, 2\pi)$. The problem is as follows. If $\alpha/2\pi$ is irrational (i.e. $\alpha \neq 2\pi k/n$ for any integers k, n), then the orbit $\{e^{in\alpha} : n \in \mathbb{Z}\}$ is dense in the circle. To see this, one can argue as follows. Let x_0 be any point on the circle and let $x_n = e^{in\alpha} x_0$. Suppose $x_n = x_m$ for some $n \neq m$. Then $e^{i(n-m)\alpha} = 1$, so $(n-m)\alpha = 2\pi k$ for some integer k, contradicting the irrationality of $\alpha/2\pi$. Hence the points x_n are all distinct. Now, if x_0 is an arbitrary point on the circle and $\varepsilon > 0$, then x_n must enter the ε-neighborhood of x_0 infinitely many times, since the circle is compact.

(d) If $\alpha/2\pi$ is rational, let $k = \min\{n > 0 : e^{in\alpha} = 1\}$ (such an integer k exists by the well-ordering principle). Show that $1, e^{i\alpha}, \ldots, e^{i(k-1)\alpha}$ are distinct and $e^{ik\alpha} = 1$.

(e) Define an equivalence relation on $\mathbb{R}$ by $x \sim y \Leftrightarrow x - y \in \mathbb{Z}$. Using the axiom of choice, form a set A by selecting one point from each distinct equivalence class. Show that A is not a Lebesgue measurable set. (If A is a Borel set with $\lambda(A) = \varepsilon > 0$, show that $\lambda(A) = 0$. Then show that $\mathbb{R}$ is a disjoint union of the sets $\{r + A : r \in \mathbb{Q}\}$. Use the translation-invariance of Lebesgue measure to show that A is not Lebesgue measurable.)

(f) Let $M = \{f : [0,1] \to [0,1] : f \text{ is Lebesgue measurable}\}$ and $M' = \{f : [0,1] \to [0,1] : f \text{ is Borel measurable}\}$. Show that $|M| = |M'| = 2^{\aleph_0}$, and any Borel subset of M or of M' has Lebesgue measure 0.

(g) Let E be an arbitrary Lebesgue measurable subset of $\mathbb{R}$. Show that $\lambda(E)$ is a Borel subset of $[0, \lambda(E)]$. (Hint: $\lambda(E) = 0$ and $\lambda(G)$ is a Borel set such that $E \subseteq G \subseteq \mathbb{R}$. Then $\mu(A) = \lambda(G \cap A)$.)

* Let Ω be a subset of $[0,1]$ with inner Lebesgue measure $\lambda_*(\Omega) = 0$ and outer Lebesgue measure $\lambda^*(\Omega) = 1$. Let $\mathcal{B}$ be the Borel subsets of $[0,1]$, μ = Lebesgue measure on $\mathcal{B}$, and μ^* = the induced outer measure. Let $\mathcal{M}$ = the μ-measurable sets. Show that $\mathcal{M} = \{B \cup N : B \in \mathcal{B}, N \subseteq \Omega\}$.

Let $\Omega = [0,1]$ and let $\mathcal{F}$ consist of all sets $(B_1 \cap \Omega) \cup (B_2 \cap \Omega^c)$, where B_1, B_2 are Borel subsets of $[0,1]$. Define

$$P((B_1 \cap \Omega) \cup (B_2 \cap \Omega^c)) = \tfrac{1}{2}(\mu(B_1) + \mu(B_2)).$$

Then $(\Omega, \mathcal{F}, P)$ is a probability space. Take $X(\omega) = \omega$, $\omega \in \Omega$.

(a) Show that P is well defined. That is, if $(B_1 \cap \Omega) \cup (B_2 \cap \Omega^c) = (B_1' \cap \Omega) \cup (B_2' \cap \Omega^c)$, then $\mu(B_1) + \mu(B_2) = \mu(B_1') + \mu(B_2')$.

(b) Suppose that $\mathcal{G} \subseteq \mathcal{F}$ is a regular conditional probability for X given $\mathcal{H} = \mathcal{B}[0,1]$. Show that $\mathcal{G}(\omega, B) = \mathcal{G}(\omega', B)$ a.s.

(c) If $H \in \mathcal{H}$, show that $\mathcal{G}(\omega, H) = 1_H(\omega)$ a.s.

(c) Show that $\mathbb{Q}(\cdot|B)$ satisfies the axioms of probability. Hint: Imitate the proof of Theorem 1.4.3.

Conclude that there is no reasonable conditional probability for P given $\emptyset$.

Note: There are many different theories of conditional probability; the one we have presented is the most popular, but there are others. For a good introduction to the subject, see Chapter 10 of Rényi (1970). For more philosophical discussion, see Shafer (1985) and the accompanying articles. For a bibliography on this material, see Section 6.10.

CHAPTER

6

STRONG LAWS OF LARGE NUMBERS AND MARTINGALE THEORY

6.1 Introduction

At the end of Chapter 1, we indicated that the classical law of convergence of the relative frequency of heads in coin tossing is best expressed as a statement about almost sure convergence of S_n/n, where S_n is a sum of independent random variables $X_1, X_2, \ldots, X_n$, or alternatively can be thought of as convergence of large numbers to such a result. Such a "strong law of large numbers" in its most general version is any statement about the almost sure convergence of a sequence of random variables, and this is the main subject matter of this chapter. A large class of convergence theorems will be developed with the aid of martingale theory, but before going into this it will be useful to consider the classical approach.

First we present some results in pure real analysis (no probability).

6.1.1 Kronecker Lemma. Let $A = \{a_n\}$ be an infinite matrix of real numbers, assume that $a_{ij} = 0$ if $j > i$ and $a_{ij} \to 0$ as $i \to \infty$ for each fixed j, and that for some nonnegative real number c, $\sum_{j=1}^{i} |a_{ij}| \le c$ for all i. If $\{x_n\}$ is a bounded sequence of real numbers, define

$$y_i = \sum_{j=1}^{i} a_{ij} x_j \qquad i = 1, 2, \ldots.$$

Then:

(a) If $x_n \to 0$ then $y_i \to 0$.

(b) If $\sum_{j=1}^{\infty} a_{ij} = 1$ and $x_n \to x$, then $y_i \to x$.

Proof. (a) We may write

$$|y_i| \le \sum_{j=1}^{N} a_{ij} |x_j| + \sum_{j=N+1}^{i} a_{ij} |x_j|. \tag{1}$$

Given $\varepsilon > 0$, choose N so that $|s_j - s| \le \varepsilon/2$ for $j \ge N$; the second term on the right-hand side of (*) is at most $\varepsilon/2$, yet $\cdots$. Since the left-hand side approaches 0 as $n \to \infty$ for any fixed N, it follows that $\sigma_n \to s$.

(ii) By (i), $\sum_{k=1}^{n} k(x_k - x_{k-1}) \to 0$, and the claim follows. $\square$

6.7.2 *Kronecker Lemma.* Let $\{a_k\}$ be a sequence of nonnegative real numbers and let $A_n = \sum_{k=1}^{n} a_k$. Assume $A_n > 0$ for all n and $A_n \to \infty$ as $n \to \infty$. If $\{x_k\}$ is a sequence of real numbers converging to the real number x, then

$$\frac{1}{A_n} \sum_{k=1}^{n} a_k x_k \to x.$$

Proof. Let a_{ij} be the infinite matrix whose ith row is

$$\left(\frac{a_1}{A_i} \quad \frac{a_2}{A_i} \quad \cdots \quad \frac{a_i}{A_i} \quad 0 \quad 0 \quad \cdots \right)$$

and apply (6.7.1). $\square$

6.7.3 *Kronecker Lemma.* Let $\{b_k\}$ be an increasing sequence of positive real numbers with $b_n \to \infty$, and let $\{y_k\}$ be a sequence of real numbers with $\sum_{k=1}^{\infty} y_k = s$ finite. Then

$$\frac{1}{b_n} \sum_{k=1}^{n} b_k y_k \to 0 \qquad \text{as} \qquad n \to \infty.$$

Proof. Let $s_n = \sum_{k=1}^{n} y_k$, with $s_0 = 0$. Then

$$\sum_{k=1}^{n} b_k y_k = \sum_{k=1}^{n} b_k (s_k - s_{k-1})$$

$$= b_n s_n - b_1 s_0 - \sum_{k=1}^{n-1} s_k (b_{k+1} - b_k).$$

Taking $b_1 = 0$. (This is the Abel summation formula; it is solved by the formula.) Thus

$$\frac{1}{b_n} \sum_{k=1}^{n} b_k y_k = s_n - \frac{1}{b_n} \sum_{k=1}^{n-1} s_k (b_{k+1} - b_k)$$

where $s_n - b_1 - b_{k+1} > 0$. Since $s_n \to s$ as $n \to \infty$ and $s_{k+1} \to s$ as $k \to \infty$,

$$= \frac{1}{b_n} \sum_{k=1}^{n-1} b_k y_k \to 0 \quad \text{as } k \to \infty. \quad \square$$

Now if S_n is a random variable with finite expectation, Chebyshev's inequality implies that

$$P\{|S_n - E(S_n)| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2} E[(S_n - E(S_n))^2] = \frac{\operatorname{Var} S_n}{\varepsilon^2}$$

If in fact S_n is a sum of independent random variables, this result can be strengthened considerably, as follows.

5.1.1 Kolmogorov's Inequality. Let $X_1, \ldots, X_n$ be independent random variables with finite expectations, and let $S_k = X_1 + \cdots + X_k$, $k = 1, \ldots, n$. Then for any $\varepsilon > 0$,

$$P\left\{\max_{1 \leq k \leq n} |S_k - E(S_k)| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2} \operatorname{Var} S_n = \frac{1}{\varepsilon^2} \sum_{k=1}^{n} \operatorname{Var} X_k$$

Proof. We may assume without loss of generality that $E(X_k) = 0$ for all k (otherwise replace X_k by $X_k - E(X_k)$). Let

$$A_k = \{|S_1| < \varepsilon, \ldots, |S_{k-1}| < \varepsilon, |S_k| \geq \varepsilon\}$$

$$A = \left\{\max_{1 \leq k \leq n} |S_k| \geq \varepsilon\right\} = \bigcup_{k=1}^{n} A_k$$

A is the disjoint union of the A_k, $k = 1, \ldots, n$. Now

$$\operatorname{Var} S_n = \int_\Omega S_n^2 \, dP \geq \int_A S_n^2 \, dP = \sum_{k=1}^{n} \int_{A_k} S_n^2 \, dP. \tag{1}$$

But $S_n = S_k + Y_k$, where $Y_k = X_{k+1} + \cdots + X_n$; hence

$$\int_{A_k} S_n^2 \, dP = \int_{A_k} S_k^2 \, dP + 2 \int_{A_k} S_k Y_k \, dP + \int_{A_k} Y_k^2 \, dP \tag{2}$$

The second term on the right-hand side of (2) is $2E[I_{A_k} S_k Y_k]$, which is 0 since $I_{A_k} S_k$ and Y_k are functions of independent random variables, and are therefore independent (see 4.6.2(c) and 4.10.8). Since the third term on the right hand side of (2) is nonnegative, we have

$$\int_{A_k} S_n^2 \, dP \geq \int_{A_k} S_k^2 \, dP \geq \varepsilon^2 P(A_k) \quad \text{by definition of } A_k$$

By (1),

$$\operatorname{Var} S_n \geq \varepsilon^2 \sum_{k=1}^{n} P(A_k) = \varepsilon^2 P(A). \quad \square$$

We shall use the Borel–Cantelli lemma (4.2.3) in quite a few of the later chapters. Here are some simple consequences which will be useful in Section 6.8 and Problem 6.8.9.

6.1.7 *Second Borel–Cantelli Lemma.* Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $A_1, A_2, \ldots$ be independent events in $\mathcal{F}$. If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.

Proof.

$$P\left(\limsup_n A_n\right) = P\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = \lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right)$$

$$= 1 - \lim_{n\to\infty} \lim_{m\to\infty} P\left(\bigcap_{k=n}^{m} A_k^c\right).$$

Now

$$P\left(\bigcap_{k=n}^{m} A_k^c\right) = P\left(\bigcap_{k=n}^{m} A_k^c\right) = \prod_{k=n}^{m} P(A_k^c) \qquad \text{by independence}$$

$$= \prod_{k=n}^{m} \exp\left[\log P(A_k^c)\right] \qquad \text{since} \qquad P(A_k^c) = \exp\left[\log P(A_k^c)\right]$$

$$= \exp\left[\sum_{k=n}^{m} \log P(A_k^c)\right]$$

$$\to 0 \qquad \text{as } m \to \infty \qquad \text{since} \qquad \sum P(A_k) = \infty.$$

Problems

1. (a) Generalize the second Borel–Cantelli lemma: Let $A_1, A_2, \ldots$ be events in a given probability space such that $\sum_{k=1}^{\infty} P(A_k) = \infty$ and

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} P(A_j \cap A_k)}{\left(\sum_{k=1}^{n} P(A_k)\right)^2} = 1.$$

(b) Show that the Borel–Cantelli zero-one law is valid. That is, if the A_n are pairwise independent and A_n are uniformly asymptotically negligible ($P(A_n) < 1$ for all n) and $\sum_{n=1}^{\infty} P(A_n) = \infty$,

(c) Use Chebyshev's inequality to show that if $\varepsilon_n = \varepsilon_0$, then

$$\liminf_{n\to\infty} P\left\{\left|\frac{S_n}{n} - m\right| \le \varepsilon_0\right\} = 1.$$

(d) Conclude from (c) that there is a sequence of integers $n_1 < n_2 < \cdots$ such that with probability 1,

$$\lim_{k\to\infty} \frac{S_{n_k}}{n_k} = m \qquad \text{for arithmetic type 1}$$

(e) Show that $P\{\lim \sup_n A_n\} = 1$.

2. Let α be uniformly distributed on $[0, 2\pi]$ and define $X_k = \sin k\theta$, $k = 1, 2, \ldots$. Show that

$$\frac{X_1 + \cdots + X_n}{n} \to 0 \qquad \text{a.e.}$$

6.2 Convergence of Random Variables

We are now in a position to establish several basic results on the convergence of sequences of random variables. We first want to describe the historical sense of the early work in this subject, the problem of certain *sums*. Let $a_1, a_2, \ldots$ be a fixed sequence of real numbers, and let ε be uniformly distributed independently over and over again. If the nth toss results in heads, we write down the number $+a_n$; if tails, the number $-a_n$. In this way we generate a series such as $a_1 - a_2 - a_3 + a_4 - \cdots$, where the signs are chosen at random. Will the series converge?

The general question suggested by the random series problem involves the convergence of sums of independent random variables. The following result gives considerable information.

6.2.1 Theorem. Let $X_1, X_2, \ldots$ be independent random variables with finite expectations. If $\sum_{n=1}^{\infty} \operatorname{Var} X_n < \infty$, then $\sum_{n=1}^{\infty} X_n - E(X_n)$ converges a.e. [The random variables are assumed to be defined on a fixed probability space $(\Omega, \mathcal{F}, P)$, and "almost everywhere" means for the probability measure P. Also, throughout this chapter, convergence will always mean as $n \to \infty$.]

Proof. We may assume that $E(X_n) = 0$ for $n = 1, 2, \ldots$. Then S_n converges if $S_n - S_k \to 0$ as $j, k \to \infty$, and this happens a.e. if for each $\varepsilon > 0$,

$$P\left[\sup_{j,k\ge n} |S_j - S_k| \ge \varepsilon\right] \to 0 \qquad \text{as} \qquad n \to \infty.$$

(see Section 2.5, Theorem 4). Equivalently, we must prove that for each $\varepsilon > 0$,

$$P\left\{\bigcup_{n=1}^{\infty}\{|S_n - S_{n-1}| \geq \varepsilon n\}\right\} \to 0 \quad \text{as} \quad m \to \infty.$$

We have

$$P\left\{\bigcup_{n=m}^{\infty}\{|S_n - S_{n-1}| \geq \varepsilon n\}\right\} = \lim_{N\to\infty} P\left\{\bigcup_{n=m}^{N}\{|S_n - S_{n-1}| \geq \varepsilon n\}\right\}$$

$$\leq \lim_{N\to\infty} P\left\{\max_{1\leq n\leq N} |S_n - S_{n-1}| \geq \varepsilon m\right\}$$

$$\leq \frac{1}{\varepsilon^2 m^2} \limsup_{N\to\infty} \operatorname{Var}(S_N - S_0) \qquad \text{by (6) of 4}$$

$$= \frac{1}{\varepsilon^2 m^2} \limsup_{N\to\infty} \sum_{n=1}^{N} \operatorname{Var}(X_n - X_{n-1}) \qquad \text{by 4.1(6)(11)}$$

$$\to 0 \quad \text{as } m \to \infty \quad \text{since} \quad \sum_{n=1}^{\infty} \operatorname{Var} X_n < \infty. \quad \square$$

To see that the alternating signs problem we have $X_1, X_2, \ldots, X_n, Y_n$, where the Y_n are independent, taking values $+1$ and -1 with equal probability. It follows that if $\sum_{n=1}^{\infty} a_n^2 < \infty$, the series $\sum_{n=1}^{\infty} X_n a_n$ converges a.e. After we prove Theorem 4 it will be clear that the condition $\sum a_n^2 < \infty$ is necessary as well as sufficient for a.e. convergence of the series.

If $X_1, X_2, \ldots$ are independent random variables we proved in Chebyshev's theorem that under appropriate conditions $(S_n - b_n)/a_n$ converges to 0 in probability (the weak law of large numbers). We now consider stronger, everywhere convergence.

4.2.2 *Kolmogorov's Strong Law of Large Numbers.* Let $X_1, X_2, \ldots$ be independent random variables with finite means and variances, and let $\{b_n\}$ be an increasing sequence of positive real numbers with $b_n \to \infty$. If

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var} X_n}{b_n^2} < \infty$$

then with $S_n = X_1 + \cdots + X_n$,

$$\frac{S_n - E(S_n)}{b_n} \to 0 \quad \text{a.e.}$$

Proof:

$$\sum_{n=1}^{\infty} \operatorname{Var}\left(\frac{X_n - E(X_n)}{n}\right) = \sum_{n=1}^{\infty} \frac{\operatorname{Var} X_n}{n^2} < \infty \qquad \text{by hypothesis.}$$

By 6.2.1, $\sum_{n=1}^{\infty} (X_n - E(X_n))/n$ converges a.e. But

$$\frac{S_n - E(S_n)}{n} = \frac{1}{b_n} \sum_{j=1}^{n} b_j \left(\frac{X_j - E(X_j)}{n}\right)$$

and this sequence converges a.e. by 6.2.3 with $b_n = n$.

In particular, if the X_n are independent random variables, each with finite mean and finite variance σ^2, then $S_n/n - \mu \to 0$ a.e. and also in L^2 (see 6.2.2).

Another special case: If the X_n are independent and the fourth central moments are uniformly bounded (that is, for some M we have $E[(X_n - E(X_n))^4] \le M$ for all n), then $\sum_{n=1}^{\infty} E[(S_n - E(S_n))/n]^4 < \infty$, so by the Cauchy–Schwarz inequality,

$$\operatorname{Var}\left(\frac{S_n - E(S_n)}{n}\right)^2 = E\left[\left(\frac{S_n - E(S_n)}{n}\right)^2\right]^2 \le E\left[\left(\frac{S_n - E(S_n)}{n}\right)^4\right] < \infty$$

and therefore 6.2.2 applies with $b_n = n$. This result due to Cantelli may in fact be proved without much machinery from measure theory, see Ash [1970], p. 208.

If the X_n are independent and have the same distribution, in other words, for each Borel set $B \subset \mathbf{R}$, $P(X_n \in B)$ is the same for all n, a version of the strong law of large numbers may be proved under a hypothesis on the mean of the X_n but no assumptions about higher moments. We first introduce some terminology that will remain in force throughout the book.

6.2.5 **Definition** If the random variables $X_1, X_2, \ldots$ have the same distribution, they will be called identically distributed. The phrase "independent and identically distributed" will be abbreviated i.i.d.

We now give a preliminary result.

6.2.6 **Lemma.** If Y is a nonnegative random variable,

$$\sum_{n=1}^{\infty} P(Y \ge n) \le E(Y) \le 1 + \sum_{n=1}^{\infty} P(Y \ge n)$$

Proof:

$$\sum_{n=1}^{\infty} P(|X_n| \geq n)$$

$$= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(k \leq |X_n| < k+1) = \sum_{k=1}^{\infty} \sum_{n=1}^{k} P(k \leq |X_n| < k+1)$$

$$= \sum_{k=1}^{\infty} k P(k \leq |X_1| < k+1) = \sum_{k=1}^{\infty} \int_{k \leq |x| < k+1} k \, dP$$

$$\leq \sum_{k=1}^{\infty} \int_{k \leq |x| < k+1} |X_1| \, dP = E(|X_1|) \leq \sum_{k=0}^{\infty} (k+1) P(k \leq |X_1| < k+1)$$

$$= \sum_{k=0}^{\infty} P(|X_1| \geq k) + \sum_{k=0}^{\infty} P(k \leq |X_1| < k+1) = \sum_{n=0}^{\infty} P(|X_1| \geq n) = 1 + E|X_1|$$

6.2.5 *Strong Law of Large Numbers, 2nd Case.* If $X_1, X_2, \ldots$ are i.i.d. random variables with finite expectation m, and $S_n = X_1 + \cdots + X_n$, then $S_n/n \to m$ a.s.

Proof: Since the X_n have the same distribution

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) = \sum_{n=1}^{\infty} P(|X_1| \geq n);$$

thus

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) \leq 1 + E|X_1| < \infty \qquad \text{by 6.2.4.}$$

By the Borel–Cantelli lemma, $P(|X_n| \geq n \text{ for infinitely many } n) = 0$. Thus if we define $Y_n = X_n$ if $|X_n| < n$; $Y_n = 0$ if $|X_n| \geq n$, then with probability 1, $Y_n = X_n$ for all sufficiently large n. Thus according to Problem 6, it is sufficient to show that

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \to m \qquad \text{a.s.}$$

Now

$$E|X_1| = E\{X_1 I_{\{|X_1| < \infty\}}\}$$

$$= E\{X_1 I_{\{|X_1| < 1\}}\} + E\{X_1 I_{\{1 \leq |X_1| < 2\}}\} + E\{X_1 I_{\{2 \leq |X_1| < 3\}}\} + \cdots$$

$$\geq 0 + 1 \cdot P(1 \leq |X_1| < 2) + 2 \cdot P(2 \leq |X_1| < 3) + \cdots$$

$$= \sum_{n=1}^{\infty} P(|X_1| \geq n) \to \infty \quad \text{as} \quad n \to \infty$$

by the Dominated Convergence Theorem.

Consequently,

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \theta_0) \to 0,$$

and therefore it is sufficient to show that

$$\frac{1}{n}\sum_{i=1}^{n} [\psi(X_i, \theta) - \psi(X_i, \theta_0)] \to 0 \quad \text{a.s.}$$

However, we know that

$$\sum_{i=1}^{n} \left| X_i - \frac{E^*(X_i)}{n} \right|$$

converges a.s., from the Kronecker lemma (1.3 with $b_n = n$ and

$$Y_n = \frac{X_n - E^*(X_n)}{n},$$

gives

$$\frac{1}{n}\sum_{i=1}^{n} [X_i - E(X_i)] \to 0 \quad \text{a.s.},$$

as desired. Now the X_i are functions of the independent random variables δ_i, and hence are independent, so by (2.3.1) it suffices to show that

$$V = \sum_{n=1}^{\infty} \mathrm{Var}\left[\frac{\psi(X_n, \theta)}{n^2}\right] < \infty.$$

(Note that $|X_i| < n$ so $\mathrm{Var}\, X_i$ is finite although nothing is known about $\mathrm{Var}\, Y_n$.) Now

$$\begin{aligned} V &= \sum_{n=1}^{\infty} \frac{1}{n^2} \mathrm{Var}\, X_n \\ &\le \sum_{n=1}^{\infty} \frac{1}{n^2} E(X_n^2) \\ &\quad \text{since } \mathrm{Var}\, X_n = E(X_n - E(X_n))^2 \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2} \psi^2(X_1, \theta) < \infty \quad \text{by the null hypothesis} \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \psi(k, \theta)^2 [F(k) - F(k-1)] \\ &= \sum_{k=1}^{\infty} \psi(k, \theta)^2 [F(k) - F(k-1)] \sum_{n=k}^{\infty} \frac{1}{n^2} \end{aligned}$$

Re-summing $\sum_i \mathbb{E}(|\eta^i|)$ with $\{\eta^i : \eta^i \neq 0\}$ we find that $\sum_1^{\infty} \mathbb{E}(|\eta^i|) \leq N\varepsilon$ for some finite positive constant N. Thus

$$\mathbb{E}\left(\sum_{i=1}^{\infty} |\eta^i|\right) \leq N\varepsilon < \infty;$$

that is, $\sum_1^{\infty} |\eta^i| < \infty$ a.s., and $\sum_1^{\infty} \eta^i = \lim_{n\to\infty} \sum_1^n \eta^i$ exists a.s. Hence

$$\mathbb{E}\left(\sum_{i=1}^{\infty} \eta^i\right) = \lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{n} \eta^i\right) = \lim_{n\to\infty} \mathbb{E}(S(X_1, \ldots, X_{n(n)})) - \mathbb{E}(S(X_1, \ldots, X_{n(n)-1})) = \mathbb{E}(X_1).$$

If $\mathbb{E}(|X_1|) = \infty$ then the above argument breaks down, but Theorem 6.3.5 can be extended to still hold. To see this, first assume that the X_i are nonnegative, with infinite expectation. If $M > 0$ and $X_i^M = \sum_{k=0}^{\infty} X_i \mathbb{I}_{[kM,(k+1)M)}(X_i)$, then all the summands are nonnegative, so

$$\liminf_{n\to\infty} \frac{S_n}{n} \geq \liminf_{n\to\infty} \frac{S_n^M}{n} = \mathbb{E}(X_1^M) \to \mathbb{E}(X_1) = \infty \quad \text{as} \quad M \to \infty.$$

Therefore $S_n/n \to \infty$ a.s. The general case is handled by splitting the random variables X_i into positive and negative parts.

We now consider the situation with a somewhat weaker result about independent random variables. If $X_1, X_2, \ldots$ are mutually independent, we might ask what is the probability that $\sum_{i=1}^{\infty} X_i$ converges? It might be expected that examples exist with any number between 0 and 1 as the probability of convergence, but in fact, the probability must be 0 or 1. More precisely, we have a definition in terms of independent random variables known as the "zero-one" property, as we shall see:

6.3.6 **Definitions and Comments.** Let $X_1, X_2, \ldots$ be a sequence of random variables, and let $\mathcal{F}_n = \sigma(X_n, X_{n+1}, \ldots)$, $n = 1, 2, \ldots$. $\mathcal{F}_n$ may be interpreted as the σ-field of events involving $X_n, X_{n+1}, \ldots$. The σ-field $\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{F}_n$ is called the *tail σ-field* of the X_n, and events in $\mathcal{T}$ are called *tail events*. Roughly speaking, tail events are those whose occurrence or nonoccurrence is not affected by changing the values of finitely many of the X_n, and a tail function is one whose value is not affected by such a change. Thus $\limsup_{n\to\infty} X_n$, $\liminf_{n\to\infty} X_n$, $\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i$, and $\{X_n \text{ converges}\}$ are tail events and tail functions, and $\limsup_{n\to\infty} X_n$ or $\liminf_{n\to\infty} X_n$ are finite-valued tail functions. [To prove a formal result, $\{\sum_{i=1}^{\infty} X_i \text{ converges}\} = \{\sum_{i=n}^{\infty} X_i \text{ converges}\} \in \mathcal{F}_n$ for each n; hence the event belongs to $\mathcal{T} = \bigcap_n \mathcal{F}_n$. Similarly,

$$\left\{\limsup_{n\to\infty} X_n < a\right\} = \left\{\limsup_{n\to\infty} X_n < a\right\} \in \mathcal{F}_n \quad \text{for all} \quad n,$$

hence $\limsup_{n\to\infty} X_n$ is $\mathcal{F}_n$-measurable.]

6.3.7 *Kolmogorov Zero–One Law.* All tail events relative to a sequence of independent random variables have probability 0 or 1, and all tail functions are constant almost everywhere.

Proof. Let $A \in \mathcal{F}_\infty$; the idea is to show that A is independent of itself, so that $P(A \cap A) = P(A)P(A)$, and consequently $P(A) = 0$ or 1. Since $\mathcal{F}_\infty \subset \mathcal{F}$, A is of the form $(X_1, X_2, \ldots) \in A'$ for some $A' \in \mathcal{B}(\mathbb{R})^\infty$. Let $\mathcal{C}$ be the class of sets $C \in \mathcal{F}(X)^\infty$ such that A and C are independent, where $C = \{(X_1, X_2, \ldots) \in C\}$. If C is a measurable cylinder, then C has the form $\{(X_1, \ldots, X_n) \in B_n\}$ where $B_n \in \mathcal{B}_{n+k}$, and we can write A in the form $\{(X_{n+1}, X_{n+2}, \ldots) \in A_{n+1}\}$, and it follows that A and C are independent. Thus $\mathcal{C}$ contains all measurable cylinders. Now if $C_1, C_2, \ldots \in \mathcal{C}$ and $C_n \uparrow C$, and $P(A \cap C_n) = P(A)P(C_n)$, $n = 1, 2, \ldots$, then $C_n \to C$ for (X_1, X_2); hence $P(A \cap C) = P(A)P(C)$. Therefore $\mathcal{C}$ is a monotone class containing the measurable cylinders; hence $\mathcal{C}$ contains all sets in $\mathcal{F}(X)^\infty$, in particular, A. But then A is independent of itself.

Finally, if f is any function other than zero on $\mathbb{R}$, then for some $c \in \mathbb{R}$, $\{\omega : f(X(\omega)) > c\}$ is a tail event, and hence has probability 0 or 1. If $A = \sup\{c \in \mathbb{R} : P\{f > c\} = 1\}$, then $f = A$ a.e. $\square$

If $\{A_n\}$ are elementary events as well as independent, a wider class of events will be shown to have probability 0 or 1.

6.3.8 *Definitions and Comments.* Let $A_1, X_2, \ldots$ be a sequence of random variables, and denote the events $\mathcal{F}_n$ as in 6.3.5. Let $A = \mathcal{F}_\infty$, so that $A = \{(X_1, X_2, \ldots) \in A'\}$, $A' \in \mathcal{B}(\mathbb{R})^\infty$. The event A is said to be *symmetric* if its nonoccurrence on finitely many coordinates of A is not affected by a permutation of finitely many of the X_n. Formally, if $\pi : \{1, 2, \ldots\} \to \{1, 2, \ldots\}$ is a permutation of finitely many coordinates, we require that $A = \{(X_{\pi(1)}, X_{\pi(2)}, \ldots) \in A\}$.

Any tail event A is symmetric, since if permutes only finitely many coordinates, we may write A in the form $\{(X_{n+1}, X_{n+2}, \ldots) \in A_{n+1}\}$ where $A \in \mathcal{B}_{n+1}$. Thus

$$A = \{(X_1, X_2, \ldots) \in \mathbb{R}^n \times A_{n+1}\} = \{(X_{\pi(1)}, X_{\pi(2)}, \ldots) \in \mathbb{R}^n \times A_{n+1}\}$$

since $\pi\{1, \ldots, n\} = \{1, \ldots, n\}$.

Here are some symmetric events that are not tail events for example $X_n = 0$ for all n, and $\{\omega : X_n(\omega) = \cdots = X_1\}$ occurs infinitely often and is less than ∞.

If $B = \{(x_1, x_2, \ldots) : (x_{\pi(1)}, x_{\pi(2)}, \ldots) \in A\}$ for each permutation π of finitely many coordinates, we denote by $N(T)$ the sequence $(X_{T(1)}, X_{T(2)}, \ldots)$ and we denote by $N(T)$ the event $\{N(T) \in B\}$. Thus A is symmetric if $A = N(T)$ for all T.

6.2.9 *Hewitt–Savage Zero–One Law.* Let $X_1, X_2, \ldots$ be iid random variables. If A is a symmetric set in $\sigma(X_1, X_2, \ldots)$, then $P(A) = 0$ or 1.

Proof. Let $\mathcal{A} = \{(X_1, X_2, \ldots) \in A\}$ so that $P^{\infty}(\mathcal{A}) = P_X(A)$. $\mathcal{A} = \{X_1, X_2, \ldots\}$. Find measurable cylinders C_n such that $P_X(A \Delta C_n) < 1/n$ where $A \Delta C_n = (A \setminus C_n) \cup (C_n \setminus A)$, and let $\mathcal{C}_n = \{A \in \mathcal{C}_n\} = \{(X_1, \ldots, X_n) \in C_n\}$. Let Γ_n interchange $(1, \ldots, n)$ and $(n+1, \ldots, 2n)$.

Since the X_i are iid, P^{∞} and $\mathcal{X}(\Gamma_n)$ have the same distribution; therefore

$$\begin{aligned} P(\mathcal{A} \Delta \Gamma_n \mathcal{C}_n) &= P_{\mathcal{X}(\Gamma_n)}(\mathcal{A} \Delta \Gamma_n \mathcal{C}_n) \\ &= P^{\infty}\{(X_{n+1}, \ldots, X_{2n}) \in A \Delta (\text{the } n\text{-cylinder } C_n)\} \\ &= P\{X \in A\} \Delta \{(X_{n+1}, \ldots, X_{2n}) \in C_n\}\} \qquad \text{since } A \text{ is symmetric} \\ &= P(A \Delta C_n) < 1/n. \end{aligned}$$

Since $P(\mathcal{A} \Delta \mathcal{C}_n) < 1/n$ and $P(\mathcal{A} \Delta \Gamma_n \mathcal{C}_n) < 1/n$, it follows from the above that $P(\mathcal{A} \Delta \Gamma_n \mathcal{C}_n) < 2/n$. It follows that $P(\mathcal{C}_n)$, $P(\Gamma_n \mathcal{C}_n)$, and $P(\mathcal{C}_n \cap \Gamma_n \mathcal{C}_n)$ all approach $P(\mathcal{A})$. But

$$\begin{aligned} P(\mathcal{C}_n \cap \Gamma_n \mathcal{C}_n) &= P\{(X_1, \ldots, X_n) \in B_n, \quad (X_{n+1}, \ldots, X_{2n}) \in B_n\} \\ &= P(\mathcal{C}_n) P(\Gamma_n \mathcal{C}_n). \end{aligned}$$

Taking $n \to \infty$ we obtain $P(\mathcal{A}) = P(\mathcal{A}) P(\mathcal{A})$, or $P(\mathcal{A}) = 0$ or 1.

Problems

1. Let $Y_1, Y_2, \ldots$ be iid random variables. If
$$\frac{1}{n} \sum_{i=1}^{n} Y_i$$
converges a.e. to a finite limit, show that $E(Y_1)$ is finite and the limit equals $E(Y_1)$ a.e.

2. The following result generalizes Lemma 6.2.4, which was used in the proof of the strong law of large numbers.

   Let F be an increasing, right continuous function on the set of non-negative real numbers to itself with $F(0) = 0$. If Y is any non-negative random variable, show that
$$\int_0^{\infty} P\{Y \geq x\} \, dF(x)$$

In particular, if $F(x) = \lambda$ we obtain

$$E[Y] = \int_0^{\infty} P[Y \geq \lambda]\, d\lambda$$

We obtain (5.24) by expressing the integral from 0 to ∞ as a sum of integrals from n to $n+1$ for $n = 0, 1, 2, \ldots$, noting that $P[Y \geq \lambda]$ is decreasing, and then applying Problem 5.1 and the Bounded Convergence Theorem. Note also that $P[Y > \lambda]$ can be replaced by $P[Y \geq \lambda]$ in the above formulas, the same proof applies.

1 Give an example to show that the "local" strong law of large numbers may fail when the X_n are uncorrelated but not identically distributed.

4 Let $X_1, X_2, \ldots$ be independent random variables, and let $g: \mathbb{R}^2 \to \mathbb{R}^1$, $h: \mathbb{R}^{n-2} \to \mathbb{R}^1$, where $n \geq 3$. If g is continuous, as other words, $g(X_1, X_2, \ldots, X_n) = g(X_1, X_2, \ldots)$ whenever Y is a permutation of finitely many coordinates, show that $g(X_1, X_2, \ldots)$ is constant a.s.

5 Let $X_1, X_2, \ldots$ be independent random variables, with $P[X_n = 1] = p_n$, $P[X_n = 0] = 1 - p_n$. Show that

$$X_n \xrightarrow{a.s.} 0 \quad \text{if} \quad \lim_{n \to \infty} p_n = 0$$

and

$$X_n \xrightarrow{a.s.} 0 \quad \text{if} \quad \sum_{n=1}^{\infty} p_n < \infty$$

6 Let $X_1, X_2, \ldots$ be nonnegative random variables, with X_n having density $\lambda_n \exp(-\lambda_n x)$, $x \geq 0$ if $\lambda_n > 0$.

(a) If $\sum_{n=1}^{\infty} \lambda_n < \infty$, show that $\sum_{n=1}^{\infty} X_n < \infty$ a.s.

(b) If the X_n are independent and $\sum_{n=1}^{\infty} \lambda_n = \infty$, show that

$$\sum_{n=1}^{\infty} X_n = \infty \quad \text{a.s.}$$

Hint: Consider $\exp(-\sum_{n=1}^{\infty} X_n)$.

7 Let $Y_1, Y_2, \ldots$ be iid random variables with $P[Y_n = 1] = P[Y_n = 0] = \frac{1}{2}$. Let $Y = (0, Y_1, Y_2, \ldots, Y_n = 1, \text{and some } x = \sum_{n=1}^{\infty} c_n 2^{-n}$. This x is in $[0, 1]$ and has a dyadic expansion $.X_1 X_2 \ldots$.

(a) Show that X is uniformly distributed on $[0, 1]$, in other words, P_X is Lebesgue measure.

(b) Show that for almost every $x \in [0, 1]$ (Lebesgue measure), the following condition holds:

For every $k = 1, 2, \ldots$ and all $j = 0, 1, \ldots, 2^k - 1$, the relative frequency of j in the k-digit dyadic expansion of x converges to $1/2^k$ as $n \to \infty$.

(c) $T = \{t \in [0,1] : X_t(\omega) = 2\}$, where ω_0 is the real number of the form $\omega_0 = 0.a_1a_2a_3\ldots$ obtained by avoiding infinite, eventually constant expansions and only a finite number of zeros. The X_t are called the Rademacher functions. The next goal is to show that $\int_0^1 \omega_0(y)\,dy = \frac{1}{2}$ and

$$\int_0^1 R_n(y)R_m(y)\,dy = \begin{cases} 0, & n \neq m, \\ 1, & n = m. \end{cases}$$

6.2 Martingales

Probability theory has its main points of interest when one is given partial information about the results in terms of a random situation. For example if $X_1, X_2, \ldots$ is a sequence of random variables, we may think of X_n as an outcome of an experiment at time n. Having observed the first n trials, one expects a fortune after trial $n+1$ is $E(X_{n+1} \mid X_1, \ldots, X_n)$. If this equals X_n, the game is "fair" since the expected gain on trial $n+1$, $E(X_{n+1} - X_n \mid X_1, \ldots, X_n) = X_n - X_n = 0$. If $E(X_{n+1} \mid X_1, \ldots, X_n) \geq X_n$, the game is "favorable" and if $E(X_{n+1} \mid X_1, \ldots, X_n) \leq X_n$, the game is "unfavorable."

We are going to study sequences of this type that arise in situations with a finite number of possible outcomes on each trial.

6.2.1 Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{\mathcal{F}_n\}_{n=1}^{\infty}$ a sequence of increasing sub-σ-fields of $\mathcal{F}$, and $\{X_n\}_{n=1}^{\infty}$ an increasing sequence of random variables on $(\Omega, \mathcal{F}, P)$. $\{X_n\}$ is said to be *adapted* to $\{\mathcal{F}_n\}$ if X_n is $\mathcal{F}_n$-measurable, that is, $X_n^{-1}(B) \in \mathcal{F}_n$ for all Borel sets $B \subset \mathbb{R}$. The sequence $\{X_n\}$ is said to be a *martingale* relative to $\{\mathcal{F}_n\}$ if (a) $\{X_n\}$ is adapted to $\{\mathcal{F}_n\}$, (b) $E|X_n| < \infty$ for $n = 1, 2, \ldots$, and (c) $E(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.e. It is a *submartingale* if (a) and (b) hold and (c)$'$ $E(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ a.e., a *supermartingale* if (a) and (b) hold and (c)$''$ $E(X_{n+1} \mid \mathcal{F}_n) \leq X_n$ a.e. The statements involving conditional expected values are "a.e." since conditional expected values are only defined almost surely.

Let $\{\mathcal{F}_n\}$ be a decreasing sequence of sub-σ-fields of $\mathcal{F}$ with X_1 essentially $\mathcal{F}_1$-measurable. If $E(X_n \mid \mathcal{F}_{n+1}) = X_{n+1}$, we say that $\{X_n, \mathcal{F}_n\}$ is a reverse martingale. Similarly, $E(X_n \mid \mathcal{F}_{n+1}) \geq X_{n+1}$ defines a reverse submartingale and $E(X_n \mid \mathcal{F}_{n+1}) \leq X_{n+1}$ defines a reverse supermartingale.

6.2.2 Theorem. (a) If $X_1, X_2, \ldots$ is a martingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) = X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

(a)$'$ If $\{X_n\}$ is a submartingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) \geq X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

(a)$''$ If $\{X_n\}$ is a supermartingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) \leq X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

(b) If $\{X_n, \mathcal{F}_n\}$ is a martingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) = X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

(b)$'$ If $\{X_n, \mathcal{F}_n\}$ is a submartingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) \geq X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

(b)$''$ If $\{X_n, \mathcal{F}_n\}$ is a supermartingale, then

$$E(X_{n+m} \mid \mathcal{F}_n) \leq X_n \quad \text{a.e.}, \quad n, m \geq 1.$$

with corresponding statements for sub- and supermartingales. For

$$\begin{aligned} E(X_{n+2} \mid \mathcal{F}_n) &= E[E(X_{n+2} \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n] \quad \text{by 5.5 (viii)} \\ &= E(X_{n+1} \mid \mathcal{F}_n) \\ &= X_n. \end{aligned}$$

The general statement follows by induction.

(c) If $\{X_1, X_2\}$ is a martingale, then

$$E(X_{n+1} \mid X_1, \ldots, X_n) = X_n \quad \text{for } n = 1, 2, \ldots$$

Thus $\{X_n\}$ is automatically a martingale relative to the sequence of fields $\sigma(X_1), \ldots, \sigma(X_1, \ldots, X_n)$. We have corresponding statements for sub- and supermartingales.

For $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ and $X_1, \ldots, X_n$ are $\mathcal{F}_i$ measurable. Since $\sigma(X_1, \ldots, X_n)$ is the smallest σ-field making $X_1, \ldots, X_n$ measurable [see 5.1.2(c)], we have $\sigma(X_1, \ldots, X_n) \subset \mathcal{F}_n$. In the martingale relation

$$E(X_{n+1} \mid \mathcal{F}_n) = X_n,$$

we take conditional expectations with respect to $\sigma(X_1, \ldots, X_n)$, we obtain the classical result by 5.8.1(h) and 5.8.1(e).

To say that (X_n) is a martingale (resp. submartingale, supermartingale) without mentioning the σ-fields $\mathcal{F}_n$, we shall always mean $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, so that $E(X_{n+1} \mid X_1, \ldots, X_n) = X_n$.

(d) $(X_1, \mathcal{F}_n)$ is a martingale if

$$\int_A X_n \, dP = \int_A X_{n+1} \, dP \quad \text{for all} \quad A \in \mathcal{F}_n, \quad n = 1, 2, \ldots$$

This follows since the condition $E(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.e. $[P]$ is equivalent to

$$\int_A E(X_{n+1} \mid \mathcal{F}_n) \, dP = \int_A X_n \, dP \quad \text{for all} \quad A \in \mathcal{F}_n$$

(see 5.8.1(a), (e)), or

$$\int_A X_{n+1} \, dP = \int_A X_n \, dP$$

by definition of conditional expectation.

Similarly, $(X_n, \mathcal{F}_n)$ is a submartingale if

$$\int_A X_n \, dP \leq \int_A X_{n+1} \, dP \quad \text{for all} \quad A \in \mathcal{F}_n,$$

and a supermartingale if

$$\int_A X_n \, dP \geq \int_A X_{n+1} \, dP \quad \text{for all} \quad A \in \mathcal{F}_n.$$

In particular, $\sigma(X_n)$ is increasing in a martingale, increasing in a submartingale, and decreasing in a supermartingale.

(d) The defining condition for a martingale relative to the orders $\sigma(X_1, \ldots, X_n)$ is equivalent to

$$E(\lambda_{n-1} X_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n) = 0 \quad \text{a.s. } P_{X_1, \ldots, X_n}$$

with similar statements for submartingales and supermartingales.

For if $A \in \sigma(X_1, \ldots, X_n)$, then A is of the form $\{(X_1, \ldots, X_n) \in B\}$, $B \in \mathcal{B}(\mathbb{R}^n)$ (see 5.4.1). If $X = (X_1, \ldots, X_n)$, then

$$\int_A Y_{n-1} \, dP = \int_{\mathbb{R}^n} Y_{n-1} \, dP_X$$

$$= \int_B E(X_n - X_{n-1} \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \, dP_X \quad \text{by 5.5.2}$$

and

$$\int_A X_n \, dP = \int_B E(X_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \, dP_X$$

$$= \int_B x_{n-1} \, dP_X \quad \text{by 5.5.1(p')}$$

The result now follows from (c).

(e) A finite sequence $\{X_k, \mathcal{F}_k, k = 1, \ldots, n\}$ is called a martingale iff $\lambda(\mathcal{F}_{k-1}, \mathcal{F}_k) = \lambda_k$, $k = 1, \ldots, n$; similar sub- and supermartingale sequences are defined similarly.

(f) If $\{X_n, \mathcal{F}_n\}$ and $\{Y_n, \mathcal{F}_n\}$ are submartingales so is $\{\max(X_n, Y_n), \mathcal{F}_n\}$.

For $E(\max(X_{n-1}, Y_{n-1}) \mid \mathcal{F}_{n-1}) \geq E(X_{n-1} \mid \mathcal{F}_{n-1}) \geq X_{n-1}$ and similarly

$$E(\max(X_{n-1}, Y_{n-1}) \mid \mathcal{F}_{n-1}) \geq Y_{n-1}.$$

This same approach shows that if $\{X_n, \mathcal{F}_n\}$ and $\{Y_n, \mathcal{F}_n\}$ are submartingales so is $\{\min(X_n, Y_n), \mathcal{F}_n\}$.

6.3.3 *Examples.* (a) If $X_n \equiv Y$ then $\{Y_n\}$ is a martingale; if $Y_n \uparrow Y$ then $\{Y_n\}$ is a submartingale; if $X_n = X_1 + \cdots$, then $\{\lambda_n\}$ is a supermartingale (sums of independent random variables with zero means).

We give some more substantial examples.

(b) Let $Y_1, Y_2, \ldots$ be independent random variables with zero means, and set $X_n = \sum_{k=1}^{n} Y_k$, $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then $\{X_n, \mathcal{F}_n\}$ is a martingale. For

$$\begin{aligned} E(X_{n-1}\mid \mathcal{F}_n) &= E(X_n + Y_{n-1} \mid Y_n, Y_{n-1}, \ldots, Y_0) \\ &= X_n - E(Y_n \mid Y_n, \ldots, Y_0) \\ &\quad \text{since} \quad X_n \text{ is } \mathcal{F}_n\text{-measurable} \\ &= X_n - E(Y_{n-1}) \\ &\quad \text{by independence (Problem 1, Section 5.5)} \\ &= X_n \\ &\quad \text{since} \quad E(Y_{n-1}) = 0. \end{aligned}$$

(b) Let $Y_1, Y_2, \ldots$ be independent random variables with $E(Y_i) = \mu_i \neq 0$ and set $X_n = \prod_{i=1}^{n} Y_i$, $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then $\{X_n, \mathcal{F}_n\}$ is a martingale, for

$$\begin{aligned} E(X_{n-1}\mid \mathcal{F}_n) &= E\left(\prod_{i=1}^{n-1} Y_i \,\middle|\, Y_n, \ldots, Y_0\right) \\ &= X_{n-1} E\left(\prod_{i=n}^{n-1} Y_i\right) \\ &\quad \text{by 5.5.1 (generalized Problem 1, Section 5.5)} \\ &= X_{n-1}. \end{aligned}$$

(c) Let Y be an integrable random variable on $(\Omega, \mathcal{F}, P)$.

If $\{\mathcal{F}_n\}$ is an increasing sequence of sub-σ-fields of $\mathcal{F}$ and $X_n = E(Y \mid \mathcal{F}_n)$ then $\{X_n, \mathcal{F}_n\}$ is a martingale. For

$$\begin{aligned} E(X_{n+1}\mid \mathcal{F}_n) &= E[E(Y \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n] \\ &= E(Y \mid \mathcal{F}_n) \qquad \text{since} \qquad \mathcal{F}_n \subset \mathcal{F}_{n+1} \\ &= X_n. \end{aligned}$$

If $\{\mathcal{F}_n\}$ is a nonincreasing sequence of sub-σ-fields of $\mathcal{F}$, and $Y_n = E(Y \mid \mathcal{F}_n)$, then $\{X_n, \mathcal{F}_n\}$ is a reverse martingale, for

$$\begin{aligned} E(X_n \mid \mathcal{F}_{n+1}) &= E[E(Y \mid \mathcal{F}_n) \mid \mathcal{F}_{n+1}] \\ &= E(Y \mid \mathcal{F}_{n+1}) \qquad \text{since} \qquad \mathcal{F}_{n+1} \subset \mathcal{F}_n \\ &= X_{n+1}. \end{aligned}$$

Note that as in 6.2.2(c), $E(X_n \mid \mathcal{F}_{n+1}) = X_{n+1}$, $n = 1, 2, \ldots$.

(d) (Branching Processes) We define a Markov chain (see 4.1) with state space $S = \{0, 1, 2, \ldots\}$. The state at time n (= number of offspring after n generations) we take $X_n = 1$, and if

$X_n = X_1 X_2 \cdots$ is the sum of a independent, identically distributed nonnegative integer-valued random variables say $Y_1, \ldots, Y_a$ where $P(Y_i = j) = a_j$, $j = 0, 1, \ldots$. Thus p_j is the probability that a given customer will spend exactly j shopping intervals; we take $p_0 = P(X_1 = 1, X_2 = 1, \ldots, X_a = 1, X_{a+1} = 0, X_{a+2} = 0, \ldots) = 1 - \sum_{j=1}^{a} p_j$.

Let $\alpha = \alpha(\theta) = \sum_{j=0}^{a} p_j e^{\theta j}$. If α is finite and θ is such that $\alpha(\theta) < 1$, then $\{\lambda_n e^{\theta n}\}$ is a martingale relative to

$$\mathcal{F}_n = \sigma(X_0, \ldots, X_n) = \sigma\left(\sum_{i=1}^{X_1} Y_i, \ldots, \sum_{i=1}^{X_n} Y_i\right).$$

For

$$E\left(\sum_{i=1}^{X_{n+1}} Y_i \mid X_n = i_n, \ldots, X_0 = i_0\right) = \sum_{j=0}^{a} p_j \sum_{m=1}^{i_n} E Y_{i_n + m}^{j}$$

(see Problem 3.3.3(a)), [1], (2) and the definition of a Markov chain)

$$= \sum_{m=1}^{i_n} \sum_{j=0}^{a} j P(Y_{i_n + m} = j)$$

$$= \sum_{m=1}^{i_n} E(Y_1 + \cdots + Y_a)$$

$$= \frac{i_n \alpha'(0)}{\alpha^0} = \frac{i_n}{\alpha^0}.$$

The result now follows from (3.3.2)(b).

(c) Consider the branching process of part (b). For $\alpha(\theta) = \sum_{j=0}^{a} p_j e^{\theta j}$, if for some $\theta > 0$ we have $\alpha(\theta) = 1$, then $\{\lambda_n e^{\theta n}\}$ is a nonnegative martingale relative to the σ-fields $\mathcal{F}(X_0, \ldots, X_n)$. For example,

$$E(\lambda_{n+1} e^{\theta(n+1)} \mid X_n = i_n, \ldots, X_0 = i_0) = \sum_{j=0}^{a} p_j e^{\theta j} \sum_{m=1}^{i_n} E Y_{i_n + m}^{j}$$

$$= \sum_{m=1}^{i_n} \sum_{j=0}^{a} p_j e^{\theta j} E(Y_1^j \mid Y_0 = i_n, \ldots, Y_{i_n} = j_{i_n})$$

$$= E(e^{\theta(Y_1 + \cdots + Y_a)}) \lambda_n e^{\theta n}$$

$$= \alpha(\theta) \lambda_n e^{\theta n} = \lambda_n e^{\theta n}.$$

If $\{Y_n, Y_n\}$ is a martingale, what can we say about Y_n or Y_n^2? In order to answer this question, we need some basic convexity theorems

6.3.4 Rate of Regret Theorem. Let $f : S \to \mathbb{R}$, where S is an open interval of reals, be convex and differentiable. Assume g is convex, that is,

$$g(\alpha x + (1-\alpha)y) \le \alpha g(x) + (1-\alpha)g(y)$$

for all $x, y \in S$ and all $\alpha \in [0,1]$. Then there are sequences $\{x_i\}$ and $\{\lambda_i\}$ of real numbers such that for all $y \in S$ we have $g(y) = \sup_i \{\alpha_i y - \lambda_i\}$.

Discussion. In the course of the proof we develop many of the basic properties of convex functions of one variable.

Take g to be a convex function from S to $\mathbb{R}$. For an open interval of reals U with $U \subseteq S$, we write λ_U for the greatest lower bound of $\{g(x) : x \in U\}$. Then by convexity,

$$g(\alpha\lambda + (1-\alpha)\lambda_U) \le \alpha g(x) + (1-\alpha)\lambda_U \le \alpha g(x) + (1-\alpha)g(y)$$

hence

$$\frac{1}{\alpha}\left[g(\alpha x + (1-\alpha)y) - g(y)\right] \le g(x) - g(y) \le \frac{1}{1-\alpha}\left[g(\alpha x + (1-\alpha)y) - g(x)\right] \tag{1}$$

and

$$-\frac{1}{\beta}[g(x-\beta) - g(x)] \le g(x) - g(x-\beta) \le \frac{1}{\beta}[g(x+\beta) - g(x)] \tag{2}$$

Also $\lambda_U \le g(x)$.

$$g(x) \le \frac{\lambda'}{\lambda'-\lambda_U} g(x) - \frac{\lambda_U}{\lambda'-\lambda_U} g(x + \lambda' - \lambda_U),$$

so

$$\frac{g(x+h) - g(x)}{h} \le \frac{g(x+\lambda'-\lambda_U) - g(x)}{\lambda'-\lambda_U}. \tag{3}$$

By (1) and (2) the right- and left-derivatives g_+' and g_-' exist on S by (3) they are equal. [Note that $x + \lambda' - \lambda_U \in (x, x+\delta)$ for some $\lambda_U < \lambda' < \lambda_U + \delta$ since S is open; this is true the difference quotients $[g(x+h) - g(x)]/h$, which decrease as $h \downarrow 0$ by (1), are bounded below by a finite constant.] Furthermore, by (1) and (3),

$$g_-'(x) = \inf_{y<x} \frac{g(x) - g(y)}{x-y}, \qquad g_-'(x) = \sup_{y<x} \frac{g(x) - g(y)}{x-y} \tag{4}$$

and by (2)

$$g_-'(x) \le g_+'(x). \tag{5}$$

If $x < z < y$, then

$$\frac{g(z)-g(x)}{z-x} \le \frac{g(y)-g(x)}{y-x} \le \frac{g(y)-g(z)}{y-z}, \tag{6.2}$$

where g' and g'' are increasing on I. The existence of right and left derivatives here implies that g has right and left derivatives at each point, and g is continuous. For if, say, g' were to jump at some point $x \in (a,b)$, then $g(x)$ would lie strictly above the line segment joining $(x-h, g(x-h))$ and $(x+k, g(x+k))$ for some $h, k > 0$. But then $g(x) \ge \alpha g(x-h) + (1-\alpha) g(x+k)$, where $x = \alpha(x-h) + (1-\alpha)(x+k)$, which contradicts (6.2). We conclude that if $g''(x) \ge 0$ on (a,b), then

$$g(x) \ge g(x_0) + g'(x_0)(x - x_0) \quad \text{for all } x, x_0 \in I. \tag{6.3}$$

The function L_{x_0} given by $L_{x_0}(x) = g(x_0) + g'(x_0)(x - x_0)$, $x \in I$, is called a *line of support* for g at the point x_0. It is immediate that $g(x) = \sup\{L_{x_0}(x) : x_0 \in I\}$ whenever $L_{x_0}(x) \le g(x)$ for all $x, x_0 \in I$, but we are trying to prove that g is the sup of essentially many lines of support. It is not hard to approximate g through a sequence of rational minorants. Thus

$$L_{x_0}(x) = g(x_0) \quad \text{for } x = x_0.$$

But $g'(x) \le g'(y) < \infty$, and the $g'(x_0)$'s, being increasing functions, are bounded on any finite closed subinterval of I, so that $L_{x_0}(x)$ forms a bounded sequence. Consequently, $L_{x_0}(x) \to g(x)$. Hence $L_{x_0}(x)$ approaches $g(x)$ through a sequence of rational minorants. That is, $g(x) = \sup g(x)$ [g is rational, $g \le L_{x_0}$ on I]. This proves (6.3), and hence the theorem. The proof is complete.

EXERCISES. 1. Show that $g(x) = \log x$ is concave on $(0, \infty)$.

Now if X is a random variable with finite mean, we have seen that $g(x) = (x-a)^2$ is convex. This is a special case of the following general convexity theorem.

6.2.3 *Jensen's Inequality.* Let g be a convex function given on an open interval I of the real line. Assume X is a random variable on $(\Omega, \mathcal{F}, P)$, with $X(\omega) \in I$ for all $\omega \in \Omega$. Suppose $E(X)$ exists as a finite or $+\infty$ or $-\infty$, where $E[g(X)]$ is defined. Then $E[g(X)] \ge g[E(X)]$ a.e. In particular, $E[g(X)] \ge g[E(X)]$.

Proof. First note that $E(X) \in I$, for if, say, $E(X) \ge b$, then $X \ge b$ a.e., and $E(X) > b$ unless

$$0 = \int_{[X \le b]} X \, dP = \int_{[X \le b]} (X - a) \, dP > 0,$$

since $X = a$ a.s. on $\{X = a\} \in \mathcal{H}$, which implies that $E[X; X = a] = a$. Thus $g(E[X \mid \mathcal{H}])$ is well-defined.

By (6.1) we may write $g(y) = \sup_n (a_n y + b_n)$, $y \in I$, on $g(I) \supset g(\mu) = \alpha$, for all n. Therefore $E[g(X)] \geq a_n E[X] + b_n$ for all n. Taking the supremum over n gives the desired result. ∎

The proof of 4.3.5 shows that the hypothesis that $E[X]$ is finite may be dropped if it is known that $E[X]$ and $E[g(X)]$ exist, and $E[X] > -\infty$.

We are now able to answer the question raised earlier about Y_t and $|X_t|$ when $\{X_t, \mathcal{H}_t\}$ is a martingale.

4.3.6 **Theorem.** (a) *Let $\{Y_t, \mathcal{H}_t\}$ be a submartingale, g a convex, increasing function from I to $\mathbb{R}$. If $g(Y_t)$ is integrable for all t, then $\{g(Y_t), \mathcal{H}_t\}$ is a submartingale. Thus, for example, if $\{Y_t\}$ is a submartingale, so is $\{Y_t^+\}$.*

(b) *Let $\{X_t, \mathcal{H}_t\}$ be a martingale, g a convex function from I to $\mathbb{R}$. If $g(X_t)$ is integrable for all t, then $\{g(X_t), \mathcal{H}_t\}$ is a submartingale. Thus if $\{X_t\}$ is a martingale and $|X_t|^p$ is integrable for all t, then $\{|X_t|^p\}$ is a submartingale.*

Proof. We have $E[g(Y_{t+1}) \mid \mathcal{H}_t] \geq g(E[Y_{t+1} \mid \mathcal{H}_t])$ by Jensen's inequality, and $E[Y_{t+1} \mid \mathcal{H}_t] \geq Y_t$, by the submartingale property, since g is increasing. Therefore $E[g(Y_{t+1}) \mid \mathcal{H}_t] \geq g(Y_t)$, since g is increasing. The proof of (b) is similar, using the martingale property, so $g(E[X_{t+1} \mid \mathcal{H}_t]) = g(X_t)$. The result follows.

Problems

1. Let $X_n = \sum_{i=1}^n Y_i$, where the Y_i are independent, with $P[Y_i = 1] = p$, $P[Y_i = -1] = q$, $p + q = 1$, $p \neq q$. Show that $\{(q/p)^{X_n}\}$ is a martingale relative to the σ-fields $\mathcal{H}_n = \sigma(X_1, \ldots, X_n) = \sigma(Y_1, \ldots, Y_n)$.

2. Consider a Markov chain whose state space is the integers, and assume that p_{ij} depends only on the difference between j and i: $p_{ij} = q_{j-i}$, where $q_n \geq 0$, $\sum_n q_n = 1$.

   (a) Show that if X_0 is the state at time 0, X_n may be written as $X_0 + Z_1 + \cdots + Z_n$, where $Z_1, Z_2, \ldots$ are independent and the Z_i all have the same distribution, namely $P[Z_i = k] = q_k$, k an integer.

   (b) If $\sum_n n q_n = 1$ (when the series is assumed to converge absolutely), show that $\{X_n^2\}$ is a martingale relative to the σ-fields $\sigma(X_0, \ldots, X_n)$.

3. Let λ be a normalized additive set function on the σ-field $\mathcal{F}$, and let $\mathcal{H}_n$ be generated by the events $A_1, A_2, \ldots, A_{k_n}$, a partition of Ω.

with $P(A_j) > 0$ for all j. Assume that the $(n+1)$st person comes into A_j so that $\mathcal{A}_n = \mathcal{A}_{n-1}$.

Define

$$X_n(\omega) = \frac{P(A_j)}{P(A_j)} \quad \text{if} \quad \omega \in A_j, \quad n = 1, 2, \ldots$$

Show that $\{X_n, \mathcal{A}_n\}$ is a martingale.

4. Define a sequence of random variables as follows: Let Z_1 be uniformly distributed between 0 and 1. Given that $Z_1 = z_1, Z_2 = z_2, \ldots, Z_{n-1} = z_{n-1}$, let Z_n be uniformly distributed between 0 and z_{n-1}. Show that $\{Z_n\}$ is a nonnegative supermartingale. Conclude that $Z_n \to 0$ a.s.

5. Let $X_1, X_2, \ldots$ be real-valued Borel measurable functions on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$. Assume that under the probability measure P on $\mathcal{B}^\infty$, $(X_1, \ldots, X_n)$ has density p_n, and under the probability measure Q on $\mathcal{B}^\infty$, $(X_1, \ldots, X_n)$ has density q_n. Define

$$Y_n = \begin{cases} \dfrac{q_n(X_1, \ldots, X_n)}{p_n(X_1, \ldots, X_n)} & \text{if the denominator is greater than } 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\{Y_n, \mathcal{F}_n(X_1, \ldots, X_n), P\}$ is a supermartingale on $(\mathbb{R}^\infty, \mathcal{B}^\infty, P)$ and $Y_n \to 0$ a.s. $[P]$ or $Y_n \to \infty$ a.s. $[P]$.

6. Let $\{Y_n, \mathcal{A}_n\}_{n \geq 0}$ be a supermartingale, and define

$$\begin{aligned} Y_0' &= Y_0 \\ Y_1' &= Y_0 - (Y_0 - E(Y_1 | \mathcal{A}_0)) \\ &\vdots \\ Y_n' &= Y_{n-1}' - (Y_{n-1}' - E(Y_n' | \mathcal{A}_{n-1})). \end{aligned}$$

Also define

$$\begin{aligned} A_0 &= 0 \\ A_1 &= Y_0 - E(Y_1 | \mathcal{A}_0) \\ &\vdots \\ A_n &= A_{n-1} + (Y_{n-1} - E(Y_n | \mathcal{A}_{n-1})) \\ &\vdots \\ A_n &= A_{n-1} + (Y_{n-1} - E(Y_n | \mathcal{A}_{n-1})). \end{aligned}$$

(a) Show that $Y_n' = Y_n - A_n$.
(b) Show that $\{Y_n', \mathcal{A}_n\}$ is a martingale.
(c) Show that $A_n \uparrow A_\infty$.

Thus a supermartingale can be expressed as the difference between a martingale and an increasing sequence. (Similarly, a submartingale can be expressed as the sum of a martingale and an increasing sequence.)

6.4 Martingales and Convergence Theorems

Under rather mild conditions, sub- and supermartingales converge almost everywhere. This result has very many consequences in probability theory.

We first prove a theorem which has an interesting gambling interpretation.

6.4.1 *Optional Skipping Theorem (Doob).* Let $X_1, X_2, \ldots$ be a submartingale (or martingale) relative to $X_1, X_2, \ldots$. Let $S_1, S_2, \ldots$ be random variables defined by

$$S_j = \begin{cases} 1 & \text{if } (X_1, \ldots, X_1) \in B_j, \\ 0 & \text{if } (X_1, \ldots, X_j) \notin B_j \end{cases}$$

where the B_j are arbitrary sets in $\mathcal{R}^j$. Define

$$\begin{aligned} Y_1 &= X_1, \\ Y_2 &= X_1 + S_1(X_2 - X_1), \\ &\vdots \\ Y_n &= X_{n-1} + S_1 S_2 \cdots S_{n-1}(X_n - X_{n-1}). \end{aligned}$$

Then $Y_1, Y_2, \ldots$ is also a submartingale and $E(Y_n) \geq E(X_n)$ for all n. If $X_1, X_2, \ldots$ is a martingale, so is $Y_1, Y_2, \ldots$, and $E(Y_n) = E(X_n)$ for all n.

Interpretation. Let X_n be the gambler's fortune after n trials; then Y_n is our fortune if we follow an optional skipping strategy. After observing $X_1, \ldots, X_{n-1}$, we must decide whether to bet with the gambler at trial n. If in this case $S_n = s_n(X_1, \ldots, X_{n-1}) = 1$, we bet and our gain $Y_n - Y_{n-1}$ is $S_1 S_2 \cdots S_{n-1}(X_n - X_{n-1})$. The theorem says that no skipping strategy can improve the game if the game is initially "fair" (a martingale) or "favorable" (a submartingale). It remains fair (favorable) and no strategy of this type can increase the expected winning.

Proof.

$$\begin{aligned} E(Y_{n+1} \mid X_1, \ldots, X_n) &= E(Y_n + S_1 \cdots S_n(X_{n+1} - X_n) \mid X_1, \ldots, X_n) \\ &= Y_n + S_1 \cdots S_n E(X_{n+1} - X_n \mid X_1, \ldots, X_n) \end{aligned}$$

since S_n is a Borel measurable function of $X_1, \ldots, X_n$ and hence is

$$\sigma(X_1, \ldots, X_n)\text{-measurable.}$$

Therefore

$$E[Y_{n+1} \mid \mathcal{F}_n] = Y_n + \nu_2 (X_n - X_n) = Y_n \qquad \text{in the martingale case,}$$

$$\geq Y_n + \nu_2 (X_n - X_n) = Y_n \qquad \text{in the submartingale case.}$$

Since $Y_1 = X_1$, we have $E[X_1] = E[Y_1]$. Arguing as above, $E[X_n - X_{n-1}] \geq 0$ ($= 0$ in the martingale case), so

$$X_n = Y_n + (X_n - X_{n-1}) + \cdots + (X_2 - X_1)$$

$$\geq Y_n + (X_{n-1} - X_{n-2}) + \cdots + (X_2 - X_1).$$

Thus

$$E[X_{n+1} - Y_{n+1} \mid \mathcal{F}_n] = (1 - \alpha_2)[X_{n+1} - X_n \mid \mathcal{F}_n] + [X_n - Y_n \mid \mathcal{F}_n]$$

$$\geq [X_n - X_{n-1} \mid \mathcal{F}_n] + \cdots + [X_2 - X_1 \mid \mathcal{F}_n]$$

with equality in the martingale case. Take expectations and use $E[E[X \mid \mathcal{F}]] = E[X]$ to obtain

$$E[X_{n+1} - Y_{n+1}] \geq E[X_n - Y_n] \geq \cdots \geq 0$$

with equality in the martingale case.

The key step in the development of the following result, due to Doob, is the inequality in Theorem 6.4(i).

6.4.2 Upcrossing Theorem Let $\{X_n, \mathcal{F}_n; n = 1, 2, \ldots\}$ be a submartingale. If a and b are real numbers, with $a < b$, let U_n be the number of upcrossings of $[a, b]$ by $X_1, \ldots, X_n$, defined as follows.

Let $T_1 = j$ (say) be the first integer in $\{1, 2, \ldots, n\}$ such that $X_j \leq a$. Let T_2 be the first integer greater than T_1 such that $X_{T_2} \geq b$. Let T_3 be the first integer greater than T_2 such that $X_{T_3} \leq a$. Define the first integer T_4 such that $X_{T_4} \geq b$, and so on. Let T_{2k+1} be the smallest integer satisfying $T_{2k} < T_{2k+1} \leq n$ and $X_{T_{2k+1}} \leq a$, if such a T_{2k+1} exists. Otherwise, let $T_{2k+1} = n$. If T_{2k} is defined and $< n$, let T_{2k+2} be the smallest integer satisfying $T_{2k+1} < T_{2k+2} \leq n$ and $X_{T_{2k+2}} \geq b$, if such a T_{2k+2} exists. Otherwise, let $T_{2k+2} = n$. Define $U_n = k$ if T_{2k} is even, and $U_n = (k-1)$ if T_{2k} is odd. Then

$$E[U_n] \leq \frac{1}{b-a} E[(X_n - a)^+].$$

Proof. First assume $a = 0$, and let $X_1 \geq 0$. Denote the T_j as above. If $X_j < a$ for every value of j, then $X_{T_1} = X_n$, but $U_n = 0$ since $T_2 = n + 1$. So $T_1 \leq j < T_2$ implies $X_j < a$. Now $T_2 \leq j < T_3$ implies $X_j \geq b$. Then $T_3 \leq j < T_4$ implies $X_j < a$, and so on. Thus by (6.4.1)

Figure A.5.1.

In Figure A.5.1, we have for this ω, $T_1 = 4$, $T_2 = 8$, $T_3 = 10$, $T_4 = 12$, $T_5 = 14$, $T_6 = \infty$, $\nu = 3$, $U_\nu = 2$,

$$\kappa_1 = \kappa_2 = \kappa_3 = 3, \qquad \kappa_4 = \kappa_5 = \kappa_6 = \kappa_7 = 4, \qquad \kappa_8 = \kappa_9 = \cdots = 6,$$

$$\kappa_{10} = 5, \qquad \kappa_{11} = \kappa_{12} = \kappa_{13} = 6, \qquad \kappa_{14} = 7,$$

$$X_1 + \sigma_1(X_2 - X_1) + \cdots + \sigma_{14}(X_{15} - X_{14}) = X_{15}$$

$$= Y_1 + (X_4 - X_3) + (X_8 - Y_4) + (X_{10} - X_9) + (X_{12} - X_{11}) + (X_{14} - X_{13}).$$

Note that Z_n, as defined in (A.5.1), is the total number of upcrossings, plus possibly a "partial upcrossing" at the end, plus a contribution due to X_1 if κ_1 is even (why?). Thus $Y_n \geq 0$. But for κ_1 as expressed in terms of $X_1, \ldots, X_n$ in the optional sampling theorem applied twice, σ_k, $k = 1, 2, \ldots, n+1$ is a submartingale, and $E(Y_n) \geq E(X_n) - E(X_1)$, i.e.

$$E(Z_n) \leq \frac{1}{a} E(|X_n|) + \frac{1}{b} E(|X_1|);$$

or, equivalently

$$E(Z_n) \leq \frac{1}{a} E(|X_n|) + \frac{1}{b} E(|X_1|).$$

In general, $\{|X_1|, \ldots, |X_n|, k = 1, 2, \ldots, n+1\}$ is a submartingale by (5.3.2), and the number of upcrossings of $[a, b]$ by $|X_i|$ is the same as the number of upcrossings of $[0, b-a]$ by $|X_i - a|^+$, provided $T_1 < \infty$. If $T_1 = \infty$, then $|X_i - a|^+ = 0$ for all i, so the upcrossings are all zero. Otherwise, as above, $Y_1 \geq 0$, $X_1 - a \geq b - a$, and $(X_1 - a)^+ \geq b - a$. The result follows from the above argument. □

We can prove the martingale convergence theorem.

A.5.3 Submartingale Convergence Theorem Let $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ be a submartingale. If $\sup_n E(X_n^+) < \infty$, then there is an integrable random variable X_∞ such that $X_n \to X_\infty$ almost everywhere.

Proof First, $X_n(\omega)$ does not converge to a finite or infinite limit if

$$P\left[\bigcup_{a<b,\, a,b \text{ rational}} \{\omega : \liminf_{n\to\infty} X_n(\omega) < a < b < \limsup_{n\to\infty} X_n(\omega)\}\right].$$

Hint: Suppose $\liminf_{n\to\infty} X_n < \alpha < \beta < \limsup_{n\to\infty} X_n$ for some α, β. Then $|X_n|$ has an infinite number of upcrossings of $[\alpha, \beta]$ on a set of positive probability, hence $P(U_{\alpha\beta}) = \infty$. But $U_{\alpha\beta}$ is the limit of the monotone sequence $U_{\alpha\beta,k}$ = the number of upcrossings of (α, β) by $X_1, \ldots, X_k$, so that $E(U_{\alpha\beta,k}) \to \infty$. But by (4.2),

$$E(U_{\alpha\beta,k}) \le (\beta - \alpha)^{-1} E(X_k - \alpha)^+ \le (\beta - \alpha)^{-1} [\sup_n E(X_n^+) + |\alpha|] < \infty$$

a contradiction. Thus X_n converges to a limit X_∞ a.e. Now $|X_n| = X_n^+ + X_n^- = 2X_n^+ - X_n$, and $E(X_n^+) \ge E(X_1^+)$ by the submartingale property. Therefore

$$E|X_\infty| \le \liminf E|X_n| \le 2 \sup_n E(X_n^+) - E(X_1) < \infty.$$

By Fatou's lemma,

$$E(|X_\infty|) \le \liminf_{m\to\infty} E(|X_m|) < \infty;$$

hence X_∞ is integrable and therefore finite a.e. By changing X_∞ on a set of measure 0, if necessary, we may take X_∞ as a random variable (rather than an extended random variable). This theorem is proved. □

(4.4) **Corollary.** *Let $\{X_1, X_2, \ldots, X_N\}$ be a finite submartingale for which $E(X_n)$ decreases as n increases, and $E(X_n, X_{n+1}) \ge X_{n+1}$ a.e. If $\inf_n E(X_n) > -\infty$, then there is an integrable random variable X_∞ such that $X_n \to X_\infty$ a.e.* [Note: For the hypothesis is satisfied for any reverse martingale since $E(X_n)$ is constant.]

Proof. Reverse the inequalities in (4.3), and let $U_{\alpha\beta,n}$ be the number of upcrossings of (α, β) by $\{X_n, X_{n-1}, \ldots, X_1\}$. Verify that this is a submartingale because:

$$E(X_n \mid X_{n+1}, \ldots, X_N) = E[E(X_n \mid X_{n+1}, \ldots, X_{N-1}) \mid X_{n+1}, \ldots, X_N] \le E(X_{n+1} \mid X_{n+1}, \ldots, X_N) = X_{n+1}.$$

We obtain $E(U_{\alpha\beta,N}) \le (\beta - \alpha)^{-1} E(X_1 - \alpha)^+ < \infty$, and thus $X_n \to X_\infty$ a.e. as before.

Now $|X_n| = 2X_n^+ - X_n$, and $E(X_n^+) \le E(X_{n+1}^+) < \infty$. Also $\{X_n^+, \ldots, X_N^+\}$ is a submartingale by (3.3)(a), so $E(X_n^+) \le E(X_N^+)$. Thus $E|X_n| \le 2E(X_N^+) - E(X_n) < \infty$, so X_∞ is integrable by Fatou's lemma. Q.E.D.

(4.5) **Comments.** (a) In (4.3) and (4.4), we proved that $\{X_n\}$ must be L^1 bounded, that is, $\sup_n E|X_n| < \infty$. This is a minimal condition, since $\sup_n E(X_n^+) < \infty$ is equivalent to L^1 boundedness, and implies convergence. However,

a submartingale (martingale) with respect to $\{\mathcal{F}_n\}$. (Compare Problems 2 and 3.)

(b) Recall Lemmas 4.3.3 and 4.4.4 and Theorem 4.4.5 for supermartingales: If $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ is a supermartingale and $\sup_n E|X_n| < \infty$, then there is an integrable random variable X_∞ such that $X_n \to X_\infty$ a.s. In particular a nonnegative supermartingale converges a.s. If $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ is a reverse supermartingale and $\sup_n E|X_n| < \infty$, there is an integrable random variable X_∞ such that $X_n \to X_\infty$ a.s. The last statement follows from the fact that $\{-X_n, \mathcal{F}_n\}$ is a submartingale and $\sup_n E(|X_n|) = \sup_n E(X_n^-)$ < ∞. (Recall from Problem 4.3.3 that $\{X_n, \mathcal{F}_n\}$ is a reverse submartingale if $X_n \geq E(X_{n-1} \mid \mathcal{F}_n)$ a.s. $= \cdots = \geq E(X_1 \mid \mathcal{F}_n)$ a.s.)

Problems

1. Consider the following Markov chain. Take $X_0 = 0$. If $X_n = 0$, regardless of $X_1, \ldots, X_{n-1}$, then

$$X_{n+1} = \begin{cases} a_{n+1} & \text{with probability} \quad p_{n+1} \\ -a_{n+1} & \text{with probability} \quad p_n \\ 0 & \text{with probability} \quad 1 - p_{n+1} - p_n \end{cases}$$

where $0 < p_{n+1} < \frac{1}{2}$ and the a_n are chosen so that $a_{n+1} > a_n$. If $X_n \neq 0$, take $X_{n+1} = X_n$. (Thus if $X_n \neq 0$, we have $X_j = X_n$ for all $j \geq n$.)

(a) Show that $\{X_n\}$ is a martingale, and X_n converges everywhere.

(b) If $\sum_{n=1}^{\infty} p_n < \infty$ or $\sum_{n=1}^{\infty} a_n p_n < \infty$, show that $\sup_n E|X_n| < \infty$.

2. (Problem 29 of W. Feller's present communication.) Consider the following Markov chain. Take $X_0 = 1$, and let

$$P\{X_n = -1 + k \mid X_{n-1} = k\} = p_n > 0,$$

$$P\{X_n = k + 1 \mid X_{n-1} = k\} = 1 - p_n > 0,$$

$$P\{X_{n+1} = 0 \mid X_n = 0\} = 1, \qquad n = 0, 1, \ldots; \quad k = 1, 2, \ldots$$

(a) Show that if $p_n = (n+1)/(2n+2)$, then all we can say about $\{X_n\}$ is that it is a martingale.

(b) If the p_n are chosen as in (a), show that $\{X_n\}$ is a martingale that converges a.s. to 0, although $E X_n \to \infty$.

(Note. In Problems 1 and 2, the Markov chain has nonstationary transition probabilities, in a sense, the impossibility of moving from state i at time n to state j at time $n+1$ depends on n. However, the basic construction of Problem 1 applies as well.)

3. (Kemeny, Snell, and Knapp.) Let $\{X_n\}$ be a Markov chain with state space S = the set of rationals in $[0,1]$, and the following transition probabilities:

   Let $0 < p < 1$, $q = 1-p$, and a be a fixed rational in $(0,1)$. If $X_n = x$, then $X_{n+1} = x+a$ with probability $p = 1-q$, and $X_{n+1} = x-(1-a)$ with probability q.

   (a) If $a = \frac{1}{2}$, show that $\{X_n\}$ is a martingale and $Y_n = X_n - X_0$ a.s., where $Y_\infty = 0$ or 1; also $P(Y_\infty = 1) = F(X_0)$.

   (b) If $a \neq \frac{1}{2}$, show that $\{X_n\}$ is a supermartingale and $X_n \to 0$ or 1.

4. (Polya's urn scheme.) An urn contains white and black balls; one ball is drawn, and then replaced by two of the same color and the process repeated. Thus if the nth selection is white and r of the $n+1$ balls in the urn before the next drawing is white, the chance of white balls in the urn before the next drawing is $(r+1)/(n+2)$.

   If X_n is the fraction of white balls in the urn just before the nth drawing, show that $\{X_n\}$ is a martingale and X_n converges a.e. to a limit X_∞, where $E(X_\infty) = E(X_1)$.

5. Martingales may be defined on a measure space $(\Omega, \mathcal{F}, \mu)$ if μ is finite, simply replacing P by the probability measure $P = \mu/\mu(\Omega)$. If $\mu(\Omega) = \infty$ we can use the following modification: $\{f_n, \mathcal{F}_n\}$ is a martingale if $f_n \in \mathcal{F}_n$ for some $\mathcal{F}_n \uparrow \mathcal{F}$ and $\int_A f_{n+1}\,d\mu = \int_A f_n\,d\mu$ for all $A \in \mathcal{F}_n$. The $\mathcal{F}_n$ are still required to be μ-measurable and integrable. Sub- and supermartingales may be defined similarly.

   Show that an L^1-bounded submartingale converges a.e. to a finite limit.

6.5 Uniform Integrability

We now introduce a concept that has important applications in martingale theory and in fact to integration theory in general.

6.5.1 Definition and Comments. Let $\{f_1, f_2, \ldots\}$ be a sequence of complex-valued, local measurable functions on the measure space $(\Omega, \mathcal{F}, \mu)$, μ σ-finite. The f_n are said to be *uniformly integrable* if

$$\int_{\{|f_n| > c\}} |f_n|\,d\mu \to 0 \quad \text{as} \quad c \to \infty$$

uniformly in n. (The definition is the same for an arbitrary index set I.) It is understood that all the f_n are locally measurable, so $|f_n|$ is integrable $\forall n$ if $|f_n| \leq g$ for all n, where g is integrable, in particular, if the f_n are uniformly bounded, then the f_n are uniformly integrable.

Unfortunately, if the f_n are uniformly integrable, then $\sup_n \int_R |f_n| dP < \infty$ necessarily if $P(R) < \infty$.

$$\int_R |f_n| dP = \int_{R \cap \{|f_n| \le M\}} |f_n| dP + \int_{R \cap \{|f_n| > M\}} |f_n| dP \le M P(R) + 1$$

for large n.

One basic application of uniform integrability is the following extension of Fatou's lemma and the Dominated convergence theorem.

6.5.2 *Theorem.* Let $f_1, f_2, \ldots$ be real valued and uniformly integrable.

(a) $\int_R \liminf_n f_n \, dP \le \liminf_n \int_R f_n dP$

$$\le \limsup_n \int_R f_n dP \le \int_R \limsup_n f_n dP$$

(b) If $f_n \to f$ in probability, then f is integrable and

$$\int_R f_n dP \to \int_R f dP.$$

Proof (a) We have

$$\int_R f_n dP = \int_{R \cap \{f_n \le -M\}} f_n dP + \int_{R \cap \{f_n > -M\}} f_n dP, \quad M > 0.$$

By uniform integrability, M may be chosen so large that $\left| \int_{R \cap \{f_n \le -M\}} f_n dP \right| < \varepsilon$ for all n, where $\varepsilon > 0$ is prespecified. Since $f_n 1_{\{f_n > -M\}} \ge -M$, which is integrable since P is finite, Fatou's lemma yields

$$\liminf_n \int_{R \cap \{f_n > -M\}} f_n dP \ge \int_R \liminf_n (f_n 1_{\{f_n > -M\}}) dP.$$

Since $f_n 1_{\{f_n > -M\}} \ge f_n$, this integral is no smaller than or equal to $\int_R \liminf_n f_n dP$. Thus

$$\liminf_n \int_R f_n dP \ge \int_R \liminf_n f_n \, dP - \varepsilon,$$

proving the $\liminf$ part. The $\limsup$ part is done by a symmetric argument.

(b) $\Rightarrow$ (a) This is immediate from (a): if $f_n \to f$ a.e., we assume $\int_\Omega |f_n - f| d\mu \not\to 0$, so for some $\varepsilon > 0$ and a subsequence f_{n_k} we have $\int_\Omega |f_{n_k} - f| d\mu \geq \varepsilon$ for all k. But then we find a subsequence $f_{n_{k_l}}$ as above, $\int_\Omega f_{n_{k_l}} d\mu \to \int_\Omega f d\mu$, a contradiction. $\square$

We now consider a useful criterion for uniform integrability.

6.5.3 **Theorem.** The complex-valued Borel measurable functions f_α are uniformly integrable if the integrals $\int_\Omega |f_\alpha| d\mu$ are uniformly bounded and also uniformly continuous, that is, $\int_E |f_\alpha| d\mu \to 0$ as $\mu(E) \to 0$ uniformly in α.

Proof: Assume the integrals are uniformly bounded and uniformly continuous. Then

$$\mu\{|f_\alpha| \geq c\} \leq \frac{1}{c} \int_\Omega |f_\alpha| d\mu$$

by Chebyshev's inequality, and the right-hand side $\to 0$ as $c \to \infty$ uniformly in α by the uniform boundedness. Thus $\int_{|f_\alpha| \geq c} |f_\alpha| d\mu \to 0$ as $c \to \infty$, uniformly in α, by the uniform continuity.

Conversely, assume uniform integrability. We have

$$\begin{aligned} \int_E |f_\alpha| d\mu &= \int_{E \cap \{|f_\alpha| \geq c\}} |f_\alpha| d\mu + \int_{E \cap \{|f_\alpha| < c\}} |f_\alpha| d\mu \\ &\leq \int_{\{|f_\alpha| \geq c\}} |f_\alpha| d\mu + c\,\mu(E). \end{aligned} \tag{1}$$

Choose c so that $\int_{\{|f_\alpha| \geq c\}} |f_\alpha| d\mu < \varepsilon/2$ for all α, if possible, and then $\mu(E) < \varepsilon/2c$ implies $\int_E |f_\alpha| d\mu < \varepsilon$. This is the uniform continuity. Uniform boundedness was verified in (1).

We have seen in 2.5.1 that L^p is a complete metric space in measure. The converse holds under certain hypotheses of uniform integrability.

6.5.4 **Theorem.** Let μ be a finite measure on (Ω, Σ), and let $0 < p < \infty$. If $f_n \xrightarrow{\mu} f$ and the $|f_n|^p$ are uniformly integrable, then $f_n \xrightarrow{L^p} f$.

Proof: The assumptions imply that $|f_n - f|^p$ are uniformly integrable. By 2.5.3, there is a subsequence f_{n_k} converging to f a.e. and in measure by 6.2.2(b),

$\int_\Omega |f_n - f|^p \, d\mu \to 0$ as $n \to \infty$. The second statement shows that any subsequence of $\{f_n\}$ has a subsequence converging to f in L^p. Hence $f_n \xrightarrow{L^p} f$. To see this, first recall that there exist $\varepsilon > 0$ and a subsequence $\{f_{n_k}\}$ such that $\int_\Omega |f_{n_k} - f|^p \, d\mu \geq \varepsilon$ for all k.

Now apply the result to the uniformly integrable sequence $g_n = |f_{n_k} - f|^p$. Since $|f|^p \leq 2^{p-1}(|f_{n_k}|^p + |f|^p)$ if $p \geq 1$, uses 2.4.b and the fact that $\{f_{n_k}\}$ is L^p-bounded, we have a subsequence $\{g_{n_{k_j}}\}$ that $\to 0$ a.e. By Egoroff's theorem, $g_{n_{k_j}}$ is integrable and it follows that the integrals $\int_\Omega |f_{n_{k_j}} - f|^p \, d\mu$ are uniformly bounded and uniformly continuous. Hence since $\int_\Omega |f_{n_{k_j}} - f|^p \, d\mu \to 0$ as $j \to \infty$ (by 2.3.5a). By 6.5.1, the $|f_{n_{k_j}} - f|^p$ are uniformly integrable, and the previous argument applies. $\square$

6.5.5 Corollary. In 6.5.4(b), $f_n \xrightarrow{L^1} f$, that is, $\int_\Omega |f_n - f| \, d\mu \to 0$.

Proof. The $|f_n|$ are uniformly integrable by hypothesis, and $f_n \xrightarrow{\mu} f$ either by hypothesis or by 2.5.5. $\square$

The following result will be useful in the next section.

6.5.6 Theorem. Let $\{f_n : n \in \mathbb{N}\}$ be integrable functions on $(\Omega, \mathcal{F}, \mu)$, μ a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose $f_n \to f$ a.e. $[\mu]$ and $\sup_{n \in \mathbb{N}} \int_\Omega |f_n| \, d\mu < \infty$. Then the f_n are uniformly integrable. For example, take $\mu(\mathbb{R}) = \infty$, $f_n = 1_{[n, n+1]}$. Then $\int_\Omega |f_n| \, d\mu \leq M$ for each n. However, the f_n are not uniformly integrable.

Proof. Given $\varepsilon > 0$, let $M = \sup_{n \in \mathbb{N}} \int_\Omega |f_n| \, d\mu$ and set $\alpha = M/\varepsilon$. There is a positive number c such that $\mu(\{x : |x| \leq c\}) > 0$ and $0 < \mu(\{x : |x| \leq c\}) < \infty$. Now

$$\int_{\{|f_n| \geq \alpha\}} |f_n| \, d\mu = \int_{\{|f_n| \geq \alpha\}} \frac{|f_n| \cdot \alpha}{\alpha} \, d\mu$$

$$\leq \int_\Omega \frac{|f_n| \cdot \alpha}{\alpha} \, d\mu = \frac{M}{\alpha} = \varepsilon. \quad \square$$

Problems

1. If $\{f_n : n \in \mathbb{N}\}$, $\{g_n : n \in \mathbb{N}\}$ are uniformly integrable and $f_n \xrightarrow{a.e.} f$, show that the $|f_n - g_n|$ are uniformly integrable.

* One can example of a uniformly integrable sequence of functions f_n on a measure space $(\Omega, \mathcal{F}, \mu)$, μ infinite, such that the $|f_n|$ cannot be majorized above by an integrable function.

6.6 Uniform Integrability and Martingale Theory

The application of the uniform integrability concept yields many additional facts about martingales. In particular, we shall find that a martingale is uniformly integrable iff it can be represented as $X_t = E(Y \mid \mathcal{F}_t)$ for some integrable random variable Y. This is the conditional analogue of the martingale $\{X_n, \mathcal{F}_n\}$

6.6.1 *Theorem.* Let P be an integrable random variable on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{F}_t$, $t \in T$, be any family of sub-σ-fields of $\mathcal{F}$. Then the random variables $X_t = E(Y \mid \mathcal{F}_t)$, $t \in T$, are uniformly integrable. That is

$$\int_{|X_t| \geq c} |X_t| \, dP \to 0 \qquad \text{as} \qquad c \to \infty$$

uniformly in t.

Proof. As $|X_t| \leq E(|Y| \mid \mathcal{F}_t)$, we have

$$\int_{|X_t| \geq c} |X_t| \, dP \leq \int_{|E(Y \mid \mathcal{F}_t)| \geq c} E(|Y| \mid \mathcal{F}_t) \, dP = \int_{|E(Y \mid \mathcal{F}_t)| \geq c} |Y| \, dP$$

since $|E(Y \mid \mathcal{F}_t)| \geq c$ if $\mathcal{F}_t$-measurable $|E(Y \mid \mathcal{F}_t)|$ is $\mathcal{F}_t$-measurable. But by Chebyshev's inequality

$$P\{|E(Y \mid \mathcal{F}_t)| \geq c\} \leq c^{-1} E(|E(Y \mid \mathcal{F}_t)|) \leq c^{-1} E(E(|Y| \mid \mathcal{F}_t)) = c^{-1} E(|Y|) \to 0$$

as $c \to \infty$ uniformly in t. ∎

The following result, due to Lévy [1937], was historically the first of the martingale convergence theorems.

6.6.2 *Theorem.* Let $\{\mathcal{F}_n\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$, and let $\mathcal{F}_\infty$ be the σ-field generated by $\bigcup_{n=1}^{\infty} \mathcal{F}_n$. If X is integrable and $X_n = E(X \mid \mathcal{F}_n)$, $n = 1, 2, \ldots$, then $X_n \to E(X \mid \mathcal{F}_\infty)$ a.e. and in L^1.

Proof. $\{X_n, \mathcal{F}_n\}$ is a martingale by 6.2(iv), and is uniformly integrable by 6.6.1. Since $E|X_n| \leq E|X| < \infty$, X_n converges a.e. to an integrable random variable X_∞ by 5.4; L^1 convergence follows from 6.3.5. It remains to show that $X_\infty = E(X \mid \mathcal{F}_\infty)$ a.e.

If $A \in \mathcal{F}_n$, then

$$\int_A X \, dP = \int_A E(X \mid \mathcal{F}_n) \, dP = \int_A X_n \, dP \to \int_A X_\infty \, dP$$

$\mathcal{F}_\infty$ measurable. Thus $\int_A X_n dP = \int_A X_\infty dP$ for all A in the field $\bigcup_{n=1}^\infty \mathcal{F}_n$ and hence for all $A \in \mathcal{F}_\infty$ (Problem 4.6.6 or the monotone class theorem). Since X_∞ is $\mathcal{F}_\infty$ measurable, X_∞ is $\mathcal{F}_\infty$ measurable and therefore $X_\infty = E(X_\infty|\mathcal{F}_\infty)$ a.s. by definition of conditional expectation. □

A martingale sometimes converges even though it is not uniformly integrable. A simple example is the following sequence.

5.5.1 Theorem Let $\{\mathcal{F}_n\}$ be a nondecreasing sequence of sub σ-fields of $\mathcal{F}$, and let $\mathcal{F}_\infty = \bigvee_{n=1}^\infty \mathcal{F}_n$. If X is integrable and $X_n = E(X|\mathcal{F}_n)$, $n = 1, 2, \ldots$, then $X_n \to E(X|\mathcal{F}_\infty)$ a.s. and in L^1.

Proof We begin by showing that $\{X_n\}$ is Cauchy in L^1; that is, $\|X_n - X_m\|_1 \to 0$ as $n, m \to \infty$, so that $\{X_n\}$ converges in L^1, so we must show that $X_\infty = E(X|\mathcal{F}_\infty)$, i.e., that $\int_A X_n dP \to \int_A X_\infty dP$ for all $A \in \mathcal{F}_\infty$, i.e., $A \in \mathcal{F}_k$ for some k. Let

$$\int_A X_n dP = \int_A E(X|\mathcal{F}_n) dP = \int_A X_k dP + \int_A X_n dP - \int_A X_k dP$$

Since X_k is $\mathcal{F}_k$ measurable for $n \geq k$, $\int_A X_k dP = \int_A X_\infty dP$ and also for $n = k$, hence $\int_A X_n dP = \int_A X_k dP$ identically. □

5.5.2 Corollary Let $\{X_n\}_{n=1}^\infty$ be a uniformly integrable martingale, and let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, then $\mathcal{F}_\infty = \sigma(X_1, X_2, \ldots)$ (see Problem 4.6.6). Thus $E(Y|\mathcal{F}_1, \ldots, \mathcal{F}_n) \to E(Y|\mathcal{F}_1, \mathcal{F}_2, \ldots)$ a.s. and in L^1. In particular, if $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, then $\mathcal{T}_n$ is the trivial σ-field $\sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots)$ (see 5.2.6).

We now show by another argument the equivalence of a submartingale's uniform integrability and L^1 convergence, and the simple case of a fact we shall soon see can be extended.

5.5.3 Theorem Let $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ be a uniformly integrable submartingale. Then $\sup_n E|X_n| < \infty$ and X_n converges to a limit X_∞ a.s. and in L^1. Furthermore, if $\mathcal{F}_\infty$ is the σ-field generated by $\bigcup_{n=1}^\infty \mathcal{F}_n$, then $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots, \infty\}$ is a submartingale, if $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ is a uniformly integrable martingale, so is $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots, \infty\}$. (If $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $n = 1, 2, \ldots, \infty$, where $\mathcal{F}_\infty$ is the σ-field generated by the $\mathcal{F}_n$, X_∞ is said to be a *last element*.)

Proof By 5.5.1, $\sup_n E|X_n| < \infty$, so by 5.4.3, $X_n \to X_\infty$ a.s. By 5.5.2, $X_n \xrightarrow{L^1} X_\infty$.

Now if $A \in \mathcal{F}_n$ and $k > n$, then $\int_A X_k dP \geq \int_A X_n dP$. Let $k \to \infty$. Since L^1 convergence yields $\int_A X_k dP \to \int_A X_\infty dP$. Thus $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots, \infty\}$ is a submartingale. The last statement follows from the above argument with "$\geq$" replaced by "$=$".

Theorem 6.6.2 indicates that uniform integrability of a submartingale implies some nice convergence properties. In fact, uniformly integrable martingales must have a last element.

6.6.4 *Theorem.* Let $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ be a uniformly integrable martingale. Then there is an integrable random variable X such that $X_n = E(X \mid \mathcal{F}_n)$, $n = 1, 2, \ldots$. In this case, $X_\infty = E(X \mid \mathcal{F}_\infty)$ exists and is the last element, where $\mathcal{F}_\infty$ is the σ-field generated by $\bigcup_{n=1}^{\infty} \mathcal{F}_n$.

Proof. The "if" part follows from Theorem 6.6.2. To prove the "only if" part, observe that $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots, \infty\}$ with $X_\infty = X$ is a martingale.

□ *Corollary.* If we require that X be $\mathcal{F}_\infty$-measurable, then X is unique (up to sets of measure 0). For if $X_n = E(X \mid \mathcal{F}_n)$, then $E(X - X' \mid \mathcal{F}_n) = E(X_n - X_n \mid \mathcal{F}_n) = 0$, $n = 1, 2, \ldots$, so $\int_A X \, dP = \int_A X' \, dP$ for all $A \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$, and hence for all $A \in \mathcal{F}_\infty$ (monotone class theorem). Thus $X = X'$ a.e. by 3.6.3.

A martingale or supermartingale with a last element need not be uniformly integrable, but we have some partial results in this direction.

6.6.5 *Theorem.* Let $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots, \infty\}$ be a nonnegative submartingale with a last element. Then the X_n are uniformly integrable.

Proof.

$$\int_{[X_n > c]} X_n \, dP \le \int_{[X_n > c]} X_\infty \, dP,$$

and

$$P[X_n > c] \le \frac{E(X_n)}{c} \le \frac{E(X_\infty)}{c} \to 0 \quad \text{as} \quad c \to \infty$$

uniformly in n. □

We now give an example of a supermartingale $\{X_n, \mathcal{F}_n\}$ with a last element that is not uniformly integrable. (Also, $\{-X_n, \mathcal{F}_n\}$ is a submartingale with a last element that is not uniformly integrable.) The key feature of this example is that $\lim_{n \to \infty} X_n$ will be a last element when the sequence is regarded as a supermartingale, but not when it is regarded as a martingale.

6.6.6 *Example.* Let $\{Y_n, n = 1, 2, \ldots\}$ be independent, with $P[Y_n = 1] = p$, $P[Y_n = 0] = 1 - p = q$, $0 < p < 1$. Let $S_n = \sum_{j=1}^{n} Y_j$, $X_n = p^{-n} \prod_{j=1}^{n} Y_j$, $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then $\{X_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ is a martingale and hence a supermartingale [see 6.3.6(iv)]. Since all $X_n \ge 0$, we have $E(X_n) = 1$ for all n. $X_\infty = 0$ is a last element when the sequence is regarded as a supermartingale, but not when it is regarded as a martingale. But the X_n are not uniformly integrable.

Since $P[Y_n = 1] = 2^{-n}$ for all n, we have $\lim_{n\to\infty} Y_n = 0$ a.s. Hence (also) $X_n = 0$ eventually, so $X_\infty = 0$ a.s. If the Y_n were uniformly integrable, then $X_\infty = 0$ by 6.6.3, since $E[Y_n] = 0$. But $E[X_n] = 1$ for all n, a contradiction. If we regard the sequence as a martingale, there can be no last element. For $E[X_\infty]$ is a fixed constant, but $X_\infty = E[Y_n | \mathcal{F}_\infty]$ is a random variable by 6.6.1, for X_n are uniformly integrable.

Since each $X_n = 0$ eventually, we have an example of a "fair" game in which the gambler is always certain to be wiped out. (Now the word "locally fair" is perhaps more appropriate than "fair.")

Note that a sub- or supermartingale with a last element converges a.e. (as the reader might easily verify for example $\sup_n E[|Y_n|] \le E[|X_\infty|] < \infty$). But it will need not converge when there is no last element.

We now look at the problem of L^p convergence.

6.6.9 Theorem. *Let $\{X_n, \mathcal{F}_n; n = 1, 2, \ldots\}$ be a martingale or a nonnegative submartingale or a nonpositive supermartingale with $\sup_n E[|X_n|^p] < \infty$ for some fixed $p > 1$. Then X_n converges to a limit X_∞ a.e. and in L^p.*

Proof. By 6.6.6, the X_n are uniformly integrable, so by 6.6.5, X_n converges a.e. to a limit X_∞, and X_∞ is a last element.

Now $\{|X_n|^p; n = 1, 2, \ldots, \infty\}$ is a nonnegative submartingale [or the case in which $\{X_n, \mathcal{F}_n\}$ is a martingale, see 6.5.6(iii)], with $\sup_{n<\infty} E[|X_n|^p] < \infty$. In the nonnegative submartingale case, use 6.6.5(ii). Write $\alpha(n) = n^{1/p}$ if $n < \infty$, $\alpha(\infty) = \lim_{n\to\infty} \alpha(n) < \infty$. By 6.6.7, the $|X_n|^p$ are uniformly integrable, and by 6.6.5, $|X_n|^p \to |X_\infty|^p$ in L^1. ∎

Problems

1. Let $\mathcal{F}_n = \mathcal{B}(X_1, \ldots, X_n)$, $Y_n = f(X_1, \ldots, X_n)$, $n = 1, 2, \ldots$, for random objects X_i in S, $\mathcal{F}_n = \sigma[Z_1, \ldots, Z_n]$, $Z_n = (X_n, Y_n)$, $n = 1, 2, \ldots$. Show that the σ-field generated by $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is $\sigma[X_1, X_2, \ldots]$.
2. Let $\{X_n, \mathcal{F}_n\}$ be a nonnegative supermartingale such that X_n converges a.e. to an integrable random variable X_∞.
   (a) Show that $E[X_n] \to E[X_\infty]$. Thus X_n are uniformly integrable. This holds for any sequence of nonnegative integrable random variables converging a.e. to an integrable limit; the supermartingale property is not involved. Note also that we have $E[|X_n - X_\infty|] \to 0$ since we have a.e. convergence and uniform integrability.
   (b) Show that X_∞ is a.e. ≥ 0.
   (c) If $E[X_n] > 0$, show that $X_\infty > 0$ a.e.

3 Let $\{X_n, \mathcal{F}_n\}$, $n = 1, 2, \ldots$, be a submartingale (respectively, a martingale) and let $\mathcal{F}_\infty$ be the σ-field generated by the $\mathcal{F}_n$. If there is an $\mathcal{F}_\infty$-measurable, integrable random variable Y such that $E(Y|\mathcal{F}_n) \geq X_n$ a.s. ($E(Y|\mathcal{F}_n) = X_n$ a.s. in the martingale case), then there is an $\mathcal{F}_\infty$-measurable, integrable random variable X_∞ such that $E(X_\infty|\mathcal{F}_n) \geq X_n$ a.s. ($E(X_\infty|\mathcal{F}_n) = X_n$ a.s. in the martingale case).

6.7 Optional Sampling Theorem

Let $X_0, X_1, \ldots$ be a martingale with X_n the present size of a gambler's total capital after n plays of a game of chance. Suppose that after each play the gambler must either stop playing or keep playing. Is it true that, in terms of quitting, what can be said about the final capital X_∞?

First of all, the random variable T must have the property that if we observe $X_0 = x_0, \ldots, X_n = x_n$, we can come to a definite decision as to whether or not $T = n$. A nonnegative integer-valued random variable of this type is called a *stopping time*.

6.7.1 *Definition.* Let $\{\mathcal{F}_n, n = 0, 1, \ldots\}$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. A *stopping time* for the $\mathcal{F}_n$ is a map $T: \Omega \to \{0, 1, \ldots, \infty\}$ such that $\{T = n\} \in \mathcal{F}_n$ for each nonnegative integer n. Since $\{T = n\} = \{T \leq n\} - \{T \leq n-1\}$ and $\{T \leq n\} = \bigcup_{k=0}^{n} \{T = k\}$, the definition is equivalent to the requirement that $\{T \leq n\} \in \mathcal{F}_n$ for all $n = 0, 1, \ldots$. If $\{\mathcal{F}_n, n = 0, 1, \ldots\}$ is a sequence of increasing σ-fields and T is a stopping time for $\{\mathcal{F}_n\}$, by definition, T is a stopping time relative to the σ-fields $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. (These always define increasing sequences.) If T is a stopping time relative to $\{\mathcal{F}_n\}$, it is a stopping time relative to $\{\mathcal{G}_n\}$ if $\mathcal{F}_n \subset \mathcal{G}_n$ for all n.

If S and T are stopping times, so are $S \vee T = \max(S, T)$ and $S \wedge T = \min(S, T)$. (Why?) $\{S \vee T \leq n\} = \{S \leq n\} \cap \{T \leq n\}$; $\{S \wedge T \leq n\} = \{S \leq n\} \cup \{T \leq n\}$. Also, if $T = c$, then T is a stopping time.

By far the most important example of a stopping time is the first time one of a set B of $\mathbb{R}^k$ is entered by a sequence $\{X_n\}$ of random variables and $B \in \mathcal{B}(\mathbb{R}^k)$. If $T(\omega) = \min\{n \geq 0 : X_n(\omega) \in B\}$ for some $\omega \in \Omega$, then since $\{T \leq n\} = \bigcup_{j=0}^{n} \{X_j \in B\} \in \sigma(X_0, \ldots, X_n)$, T is a stopping time since $\{T = n\} = \{X_0 \notin B, \ldots, X_{n-1} \notin B, X_n \in B\}$.

If T is a stopping time for $\{X_n\}$, an event A is said to be "prior to T" if, whenever $n = 1, 2, \ldots$, we can tell by examination of the $X_0, \ldots, X_n$ whether or not A has occurred. The formal definition is as follows.

6.7.2 *Definition.* Let T be a stopping time for the σ-fields $\mathcal{F}_n$, $n = 0, 1, \ldots$, and let A belong to $\mathcal{F}$. The set A is said to be prior to T if $A \cap \{T \leq n\} \in \mathcal{F}_n$ for all $n = 0, 1, \ldots$. (Equivalently, as in 6.7.1, $A \cap \{T = n\} \in \mathcal{F}_n$ for all $n = 0, 1, \ldots$.) The collection of all sets prior to T will be denoted by $\mathcal{F}_T$. It follows quickly that $\mathcal{F}_T$ is a σ-field. Also, if $T = n$, then $\mathcal{F}_T = \mathcal{F}_n$.

If S and T are stopping times and $S \le T$, then $\mathcal{F}_S \subset \mathcal{F}_T$. For if $A \in \mathcal{F}_S$ then

$$A \cap \{T \le t\} = \bigcup_{n=0}^{\infty} \{n \le S \le t\} \cap A \cap \{T \le t\}.$$

But $A \cap \{n \le S \le t\} \in \mathcal{F}_t$ and $\{T \le t\} \in \mathcal{F}_t$, hence $A \in \mathcal{F}_T$.

If the stopping time T is constant, $T \equiv n$, then $\mathcal{F}_T = \mathcal{F}_n$—no surprise. We would like this idea to carry over to a general stopping time. Naturally, let T be a finite stopping time. To the σ-fields $\mathcal{F}_n$ and define $\mathcal{F}_T$ in the natural way: if $T(\omega) = n$, let $X_T(\omega) = X_n(\omega)$. Then $\{T = n\} \in \mathcal{F}_n$, hence $\{X_T \in B\} \in \mathcal{F}_n$, in other words X_T is $\mathcal{F}_T$ measurable. (Since $\mathcal{F}_T \subset \mathcal{F}$ by definition, it follows in particular that X_T is a random variable.) To see this, write

$$\{X_T \in B\} \cap \{T = n\} = \bigcup_{k=0}^{n} \{X_k \in B\} \cap \{T = k\}.$$

Since $\{X_k \in B\} \in \mathcal{F}_k \subset \mathcal{F}_n$ and $\{T = k\} \in \mathcal{F}_n$, we have

$$\{X_T \in B\} \cap \{T = n\} \in \mathcal{F}_n.$$

Also, as T is finite, we have $\bigcup_{n=0}^{\infty} \{T = n\} = \Omega$, so that $\{X_T \in B\} \in \mathcal{F}_T$ as claimed.

If T is not necessarily finite, the same argument shows that $X_T \cdot 1_{\{T < \infty\}}$ is $\mathcal{F}_T$ measurable.

Now in the gambling situation described at the beginning of the section, a basic quantity of interest is $E(X_T)$: the average accumulation at the stopping time. For example, $E(X_\tau)$ turns out to be the same as $E(X_1) = E(X_0)$: On all of the meanings of gambling, the gambler's strategy does not affect any improvement over the proportion of winning games. Now in computing X_τ and X_T we are considering two stopping times S and T with $S \le T$, and looking at X_T versus X_S. More generally, if $T_1 \le T_2 \le \cdots$ form an increasing sequence of finite stopping times, we may examine the sequence $X_{T_1}, X_{T_2}, \ldots$. If the sequence forms a martingale, then $E(X_{T_2}) = E(X_{T_1})$. So, e.g., if $T = 1$, then $E(X_T) = E(X_1)$.

Thus if we interpret the gambler's tomorrows as random times $T_1 < T_2 < \cdots$, the basic question is whether the martingale (or submartingale) property is preserved. This will always be the case when the sequence $\{X_n\}$ is finite.

6.2.7 *Theorem.* Let $\{X_n, \mathcal{F}_n\}_{n=0}^{N}$ be a submartingale, and let $T_1, T_2, \ldots, T_N$ be an increasing sequence of stopping times for the $\mathcal{F}_n$. In other

suppose that T_1 takes values in $\{1, \ldots, n\}$ and $\{T_1 \le k\} \in \mathcal{F}_k$, $k = 1, \ldots, n$. The σ-fields $\mathcal{F}_{T_1}$ are defined as before:

$$\mathcal{F}_{T_1} = \{A : A \cap \{T_1 \le k\} \in \mathcal{F}_k, \ k = 1, \ldots, n\}.$$

Then the X_{T_1} form a submartingale relative to the σ-fields $\mathcal{F}_{T_1}$, a martingale if $\{Z_k\}$ is a martingale.

THEOREM. (Hoffman-Jørgensen (1968)). Define $Y_k = X_{T_k}$, and note that each Y_k is integrable.

$$\int_A |Y_k| \, dP = \sum_{j=1}^{n} \int_{A \cap \{T_k = j\}} |X_j| \, dP \le \sum_{j=1}^{n} E|X_j| < \infty$$

As the T_k increase with k, so do the $\mathcal{F}_{T_k}$ form the filtration after the T_k.

Now if $A \in \mathcal{F}_{T_{k+1}}$, we must show that $\int_A Y_{k+1} \, dP \ge \int_A Y_k \, dP$ with equality in the martingale case. Since $A = \bigcup_j A \cap \{T_{k+1} = j\}$, it suffices to replace A by $A_j = A \cap \{T_{k+1} = j\}$, which belongs to $\mathcal{F}_j$ now. If $j > 1$, we note that $T_{k+1} = j$ implies $T_k = 1, \ldots, j$, so that

$$\int_{A_j} Y_{k+1} \, dP = \sum_{i=1}^{j} \int_{A_j \cap \{T_k = i\}} X_{T_{k+1}} \, dP + \int_{A_j \cap \{T_k > j\}} X_j \, dP$$

Thus

$$\int_{A_j} Y_{k+1} \, dP = \sum_{i=1}^{j-1} \int_{A_j \cap \{T_k = i\}} X_i \, dP - \int_{A_j \cap \{T_k > j\}} X_j \, dP$$

$$+ \int_{A_j \cap \{T_k \ge j\}} (X_j - Y_{k+1}) \, dP \qquad (*)$$

Now consider the $i = j$ term in (*) with the $\int X_j \, dP$ next to obtain

$$\int_{A_j \cap \{T_k \le j\}} X_j \, dP - \int_{A_j \cap \{T_k = j\}} X_j \, dP = \int_{A_j \cap \{T_k < j\}} X_j \, dP$$

$$\le \int_{A_j \cap \{T_k < j\}} X_{j-1} \, dP$$

since $X_{j-1} \le X_j$ and $A_j \cap \{T_k < j\} \in \mathcal{F}_{j-1}$. But

$$\int_{A_j \cap \{T_k < j\}} X_{j-1} \, dP = \int_{A_j \cap \{T_k \le j-1\}} X_{T_k} \, dP,$$

so the term can be combined with the $i = 1, \ldots, j-1$ terms of (*) to obtain

$$\int_{A_j \cap \{T_k \le j-1\}} X_{T_k} \, dP.$$

Proceeding inductively we find

$$\int_{D_n} X_{n+1}\,dP \le \int_{D_n \cap [\tau_n = n]} X_n\,dP + \int_{D_n \cap [\tau_n \ge n+1]} X_{n+1}\,dP. \qquad (2)$$

Now $D_n \cap [\tau_n \ge n+1] = D_n \cap [\tau_n \ge n]^c$ is simply D_n if $\tau_n \ge n$. Finally, $D_n \cap [\tau_n = n] = D_n - D_{n+1}$ since $D_{n+1} = [\tau_n = n]^c$ and $D_n = D_{n+1}$ on D_n. Thus

$$\int_{D_n} X_{n+1}\,dP \le \int_{D_n} X_n\,dP$$

as desired. In the martingale case all inequalities in the proof become equalities. □

Theorem 3.7.2 extends immediately to the case of an infinite sequence $\{X_n\}$ such that τ_n is bounded, that is, for each n there is a positive constant K_n such that $\tau_n \le K_n$ a.e. The same proof may be used; the key point is that $P[\tau_n \ge k]$ is still strictly less than 1 for sufficiently large k.

When $\{X_n\}$ is an infinite sequence, the submartingale or submartingale property is not preserved in general, but the following result gives several sufficient conditions.

6.7.4 *Optional Sampling Theorem.* Let $\{X_1, X_2, \cdots\}$ be a submartingale, and let $\tau_1, \tau_2, \cdots$ be an increasing sequence of finite stopping times for $\{X_n\}$, with $\tau_n = \tau_{n+1} = n$, $n = 1, 2, \cdots$. If

(A) $E|X_{\tau_n}| < \infty$ for all n, and

(B) $\liminf_{k\to\infty} \int_{[\tau_n \ge k]} |X_k|\,dP = 0$ for all n,

then $\{X_{\tau_n}\}$ is a submartingale relative to the new fields $\mathcal{F}_{\tau_n}$. If $\{X_n\}$ is a martingale, so is $\{X_{\tau_n}\}$.

Proof. Since integrability of the X_{τ_n} is now hypothesis (A), we can follow the proof of 6.7.2 to (2). The first integral on the right is $\int_{D_n} X_n\,dP$ as before, but in the second integral we no longer have $P[\tau_n \ge n] = 0$ simply for large k. But by hypothesis (B), $\int_{[\tau_n \ge k]} |X_k|\,dP < \varepsilon$ if $k > n_0$ for a given $\varepsilon > 0$ and n. Proceeding through an appropriate subsequence, we have $\int_{D_n \cap [\tau_n \ge k]} |X_{k+1}|\,dP < \varepsilon$ for $k > n_0$ since $P[\tau_n \ge k]$ decreases to the empty set. Thus

$$\int_{D_n} X_{\tau_n}\,dP \le \int_{D_n} X_n\,dP$$

as desired. As before, all inequalities become equalities in the martingale case. □

If X_{∞} is a submartingale and a last element X_{∞} is adjoined, then we can define the random variable X_T for any stopping time T. On the set $[T = \infty]$, $X_T = X_{\infty}$. In this case, the optional sampling theorem holds.

6.7.5 *Theorem.* Suppose $\{T_n\}$ is an increasing sequence of stopping times (not necessarily finite) for a submartingale $\{X_n\}$ having a last element X_{∞}. Then $\{X_{T_n}\}$ is a submartingale relative to the sequence $\mathcal{B}_{T_n}$; $\{|X_n|\}$ is a submartingale, so is $\{|X_{T_n}|\}$. In particular, this holds if the X_n are uniformly integrable (see 6.5.1).

Proof. Suppose $T_1 \leq T_2$. If $A \in \mathcal{B}_{T_1}$, then $A \cap [T_1 = n] \in \mathcal{B}_n$ for every finite n, so $A \cap [T_2 = n] = A \cap [T_1 = n] \cap [T_2 = n] \in \mathcal{B}_n$. It follows that $A \in \mathcal{B}_{T_2}$, so $\mathcal{B}_{T_1} \subset \mathcal{B}_{T_2}$. Now if T is a stopping time, $T \wedge n = \min(T, n)$ is easily checked to be a stopping time for the $\mathcal{B}_n$. Now if T_n is finite, $X_{T_n} = X_{T_n \wedge n} + \sum_{k=1}^{n} (X_k - X_{k-1}) I_{[T_n \geq k]}$ for $n = 1, 2, \ldots$. Hence by Doob's theorem,

$$\limsup_{n \to \infty} \int_{[T_n < \infty]} X_{T_n} \, dP = \limsup_{n \to \infty} \int_{[T_n < \infty]} X_{T_n \wedge n} \, dP \leq \int_{\Omega} |X_1| \, dP < \infty.$$

The same conclusion holds for arbitrary T_n, because on the set $[T_n = \infty]$ we have $X_{T_n} = X_{\infty} \geq 0 = X_{T_n}$. But by 6.7.3, $\{|X_n|\}$ is a submartingale; hence

$$\int_A X_{T_n} \, dP \geq \int_A X_{T_{n+1}} \, dP = \int_A X_{T_n} \, dP$$

which is finite because X_{T_n} is integrable.

Again by 6.7.3, $\{X_{T_n \wedge k}, \mathcal{B}_{T_n \wedge k}, k = 1, 2, \ldots\}$ is a submartingale. Hence we have

$$\int_A X_{T_n \wedge k} \, dP \leq \int_A X_{T_{n+1} \wedge k} \, dP$$

if $A \in \mathcal{B}_{T_n \wedge k}$. But if $A \in \mathcal{B}_{T_n}$, then $A \cap [T_n \leq k] \in \mathcal{B}_{T_n \wedge k}$, for

$$A \cap [T_n \leq k] \cap [T_n \wedge k = i] = \begin{cases} A \cap [T_n = i] & \text{for} \quad i < k, \\ A \cap [T_n \geq k] & \text{for} \quad i = k. \end{cases}$$

Thus

$$\int_{A \cap [T_n \leq k]} X_{T_n \wedge k} \, dP \leq \int_{A \cap [T_n \leq k]} X_{T_{n+1} \wedge k} \, dP \qquad A \in \mathcal{B}_{T_n}.$$

But on $[T_n \leq k]$, $T_n \wedge k = T_n$, also $[T_{n+1} \leq k] \subset [T_n \leq k]$ and $X_{T_{n+1} \wedge k} \geq 0$, hence

$$\int_{A \cap [T_n < \infty]} X_{T_n} \, dP \leq \int_{A \cap [T_{n+1} < \infty]} X_{T_{n+1}} \, dP$$

Let $A \in \mathcal{F}_{S \wedge T}$. Then

$$\int_A X_{S \wedge T} \, dP = \int_{A \cap \{S \le T\}} X_{S \wedge T} \, dP + \int_{A \cap \{S > T\}} X_{S \wedge T} \, dP.$$

As $X_{S \wedge T} = X_S$ on $\{S \le T\}$ and $X_{S \wedge T} = X_T$ on $\{S > T\}$, we get

$$\int_A X_{S \wedge T} \, dP \le \int_A X_S \, dP + \int_A X_T \, dP,$$

which is the desired submartingale property.

(Case 1) $Z_k = E(X_{kn} | \mathcal{F}_{(k-1)n})$, $k = 1, 2, \ldots$, are the steps $X_{jn} - X_{(j-1)n}$ in the original submartingale. Hence $E(Z_k Y_{(k-1)n}) = E(X_{kn} Y_{(k-1)n})$ by (6.1.1) and 6.3.10(a). Therefore, the Z_k, $k = 1, 2, \ldots$, are an independent sequence by (6.6.1).

Now if $A \in \mathcal{F}_S$, then $A \cap \{T_j = n\} \in \mathcal{F}_{jn}$ for each n. Thus

$$\int_{A \cap \{T_j = n\}} Y_{(j-1)n} \, dP = \int_{A \cap \{T_j = n\}} X_{jn} \, dP.$$

Let $n \to \infty$ and use the uniform integrability of the $Y_{(j-1)n}$ to obtain

$$\int_{A \cap \{T_j < \infty\}} Y_{j-1} \, dP = \int_{A \cap \{T_j < \infty\}} X_{T_j} \, dP.$$

But on $\{T_j = \infty\}$ we have $X_{T_j} = X_\infty = 0$, so

$$\int_A Y_{j-1} \, dP = \int_A X_{T_j} \, dP \qquad \text{for every} \qquad A \in \mathcal{F}_S.$$

Therefore $Y_{j-1} = E(X_{T_j} | \mathcal{F}_S)$ so that $Y_0, Y_1, \ldots$ is a martingale.

(General Case) We let $X_n = Z_n - M_n$, where $M_n = E(Z_\infty | \mathcal{F}_n)$. $Y_n = Z_{T_n} - M_{T_n}$. The Y_j is independent of the $\mathcal{F}_0$ and the proof follows. Note that $E(X_n)$ is a martingale with last element X_∞, we must have $X_n = E(Y_n | \mathcal{F}_n)$ so that (Y_n) is a martingale by the analysis of case 2. $\square$

In contrast to the section we give an example of a situation in which the optional sampling theorem does not apply. Consider the problem of coin tossing, that is, let $Y_1, Y_2, \ldots$ be independent random variables each taking on values $+1$ and -1 with equal probability. If $X_n = Y_1 + \cdots + Y_n$ or X_n forms a martingale by (6.1.2a). Now with probability 1, $X_n = 1$ for some n. (This is a standard random walk result. For a proof, see Ash, 1970, p. 288.) If T is the time that 1 is reached, (or Tchebycheff's inequality), and $S = 0$, then S and T are each finite stopping times, but $X_S = 0$, $X_T = 1$, a contradiction. In addition, we obtain from (6.3.1) the result that the X_n are not uniformly integrable.

Problems

1. Let $\{Z_1, \ldots, Z_n\}$ be a submartingale and let T be a stopping time for $\{X_1, \ldots, X_n\}$. Show that

$$E(Z_T) \le E(Z_{T+1}) \le \cdots \le E(Z_n).$$

   The corresponding result for supermartingales, which may be obtained by replacing X_i by $-X_i$, is

$$E(Z_T) \ge E(Z_{T+1}) \ge \cdots \ge E(Z_n).$$

2. Let $\{X_1, X_2, \ldots\}$ be a submartingale, and T a finite stopping time for $\{X_i\}$. Show that

$$E(|X_T|) \le 2 \sup_i E(X_i^+) - E(X_1).$$

   As in Problem 1, we may also write down the supermartingale version:

$$E(|X_T|) \ge 2 \inf_i E(X_i^-) + E(X_1).$$

3. (Wald and supermartingale inequalities) (a) Let $\{X_1, \ldots, X_n\}$ be a submartingale. If $\lambda \ge 0$, show that

$$\lambda P\left\{\max_{1 \le i \le n} X_i \ge \lambda\right\} \le \int_{\{\max X_i \ge \lambda\}} X_n \, dP \le E|X_n|.$$

   (b) Let $\{X_1, \ldots, X_n\}$ be a submartingale. If $\lambda \ge 0$, show that

$$\lambda P\left\{\min_{1 \le i \le n} X_i \le -\lambda\right\} \le F(X_1) - \int_{\{\min X_i \le -\lambda\}} X_n \, dP$$

$$\le F(X_1) + \lambda E(X_n^-).$$

   (Apply 6.2.1 with $Y = \min\{X_i, -\lambda\}$. $F = \alpha$ if all $X_i > -\lambda$.)

   (c) If $\{X_1, X_2, \ldots\}$ is a submartingale and $\lambda > 0$, show that

$$\lambda P\{\sup_i X_i \ge \lambda\} \le \sup_i E(X_i^+);$$

   if $\{X_1, X_2, \ldots\}$ is a supermartingale and $\lambda > 0$, show that

$$\lambda P\{\sup_i X_i \ge \lambda\} \le E(X_1) - \inf_i E(X_i^-).$$

4. Use Problem 3 to give an alternate proof of Kolmogorov's inequality 6.2.1.

5 (Wald's equation for the case of a random number of random variables) Let $X_1, X_2, \ldots$ be independent, identically distributed, absolutely integrable random variables with finite mean m, and let $S_n = \sum_{i=1}^{n} X_i$. If T is a finite stopping time for $\{X_n\}$ which satisfies the following:

(a) $E(X_n I_{\{T \geq n\}}) = E(X_n)P(T \geq n)$ for all $n \geq 1$,

(b) $\sum_{n=1}^{\infty} E(|X_n| I_{\{T \geq n\}}) < \infty$,

[Recall that $I_A = 1$ if A occurs and equals 0 if A does not occur. Use (6.1.5) to prove (a).]

(c) T is a general (i.e., not necessarily integer-valued) random variable that is independent of $\{X_1, X_2, \ldots\}$ but not necessarily a stopping time, show that the result may not be still true.

6 [Absorption probability for the simple moving average (Moran, 1958)] Let

$$Y_n = \frac{1}{2}(X_n + X_{n-1}) \quad (n = 1, 2, \ldots),$$

be a moving average, where $X_0, X_1, \ldots$ are independent random variables each taking the values $+1$ and -1 with probability $\frac{1}{2}$ each. Define the stopping times T_1 as in (6.1.3) and the corresponding set $S_n = Y_{T_n}$.

(a) Show that $S_n = \frac{1}{2}(X_0 + X_{T_n})$ (take $X_{T_n} = 0$).

(b) Show that $E(T_1) = 2E(T_0)$; the general corresponding difference equation is then calculated just as in (6.4.2). [Since $E(X_{T_{n-1}+1} \mid X_{T_{n-1}}) = 0$ one has $X_{T_n} = X_{T_{n-1}}$ with probability $\frac{1}{2}$ if $T_n < \infty$ and $T_n = \infty$ with probability $\frac{1}{2}$ if $T_n < \infty$.]

$$E(X_0^2) = \sum_{n=1}^{\infty} E(X_{T_n}^2 - X_{T_{n-1}}^2) = \frac{1}{2}E(T_1).$$

6.8 Applications to Martingale Theory

Martingale ideas provide fresh insights and simplifications for many problems in probability. In this section we consider some important examples. First, we use the martingale convergence theorem to provide a short proof of the strong law of large numbers for sums of random variables (see 6.2.5). We need two preliminary facts.

6.8.1 Lemma. *Let $X_1, \ldots, X_n$ be pairwise independent identically distributed random variables with finite expectation, and $S_n = \sum_{i=1}^{n} X_i$. Then*

$$E(X_k \mid S_n) = \frac{S_n}{n} \quad \text{a.s.} \qquad (k = 1, \ldots, n).$$

Intuitively, given $S_n = X_1 + \cdots + X_n$, the average contribution of each X_k is the same, and hence must be S_n/n.

6.3.1 Theorem. If $X_1, X_2, \ldots$ are i.i.d. and $E|X_1| < \infty$, then

$$\int_{[S_n \in B]} X_k \, dP = \int_{[S_n \in B]} S_n/n \, dP$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1 I_B(x_1 + \cdots + x_n) \, dF(x_1) \cdots dF(x_n)$$

where F is the distribution function of the X_i. By Fubini's theorem, this is independent of k, hence

$$\int_{[S_n \in B]} X_k \, dP = \frac{1}{n} \sum_{i=1}^{n} \int_{[S_n \in B]} X_i \, dP = \int_{[S_n \in B]} \frac{S_n}{n} \, dP. \quad \square$$

6.3.2 Lemma. If $X_1, X_2, \ldots$ are random variables and $S_n = \sum_{i=1}^{n} X_i$, then $\sigma(S_n, X_{n+1}, X_{n+2}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$.

Proof. Since $X_{n+1} = S_{n+1} - S_n$, $X_{n+2} = S_{n+2} - S_{n+1}$, $\ldots$, the left-hand σ-field $\subset \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$. Conversely, $S_{n+1} = S_n + X_{n+1}$, $S_{n+2} = S_n + X_{n+1} + X_{n+2}$, $\ldots$. Hence $\sigma(S_n, X_{n+1}, X_{n+2}, \ldots) \subset \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$. $\square$

6.3.3 Strong Law of Large Numbers, 1st Case. If $X_1, X_2, \ldots$ are i.i.d. random variables with finite expectation m and $S_n = X_1 + \cdots + X_n$, then $S_n/n \to m$ a.e. and in L^1.

Proof. $(X_1, S_n, S_{n+1}, S_{n+2}, \ldots)$ and $(X_2, S_n, S_{n+1}, S_{n+2}, \ldots)$ are independent [Problem Sect. 5.4, Problem 11]; hence (X_1, S_n) and (X_2, S_n) are identically distributed independent random objects, and independently $\sigma(S_n)$. Therefore

$$E(X_1 | S_n) = E(X_2 | S_n, S_{n+1}, S_{n+2}, \ldots) \quad \text{by Problem 11, Section 5.3}$$

$$= E(X_1 | S_n, S_{n+1}, S_{n+2}, \ldots) \quad \text{by 6.3.2.}$$

Thus by 6.3.1,

$$E(X_1 | S_n, S_{n+1}, \ldots) = \frac{S_n}{n} \quad \text{a.e.}$$

But by 6.3.2 and 6.2.4, $E(X_1 | S_n, S_{n+1}, \ldots) \to E(X_1 | \mathcal{T})$ a.e. and in L^1, where $\mathcal{T} = \bigcap_n \sigma(S_n, S_{n+1}, \ldots)$ is the tail σ-field of the X_i. Since S_n/n is constant a.e. and in L^1 as a limit, $S_n/n \to m$ a.e. and in L^1.

Now to show that $S_n/n \to 0$ a.s. we may proceed in two ways. One approach is to note that $\lim_{n\to\infty} |X_n|/n = 0$ a.s. by the Kolmogorov zero-one law 2.2.3. Since S_n/n is L^1 convergent and $E|S_n/n| \to \infty$, the constant must be ∞.

Alternatively, we may use the Hewitt–Savage zero-one law 5.2.3 to show that each a in $\mathcal{S}_\infty$ has probability 0 or 1. For if $A = \mathcal{S}_\infty$ and T permutes coordinates, then $A = m\{S_n \circ \theta_{n+1} \cdots\}$ hence A is the same as $[S_n \circ \theta_{n+1} \cdots \in B]$ for some $B \subset \mathcal{B}(\mathbb{R}^n)$. Since $S_n = X_1 + \cdots + X_n = Y_{n+1} + \cdots$ $Y_{n,n} = X_{n+1} + \cdots + X_{2n}$ is symmetric, and, therefore, the Hewitt–Savage zero-one law is applicable. Thus $P(A) = 0$ or 1, and it follows that $E(X_1 Y_{n,n}) = 0 (= E(X_1)E(Y_{n,n}))$ by 5.5.7(iii). ∎

Notice that we have obtained L^1 convergence in the strong law of large numbers. This would be more difficult to derive using the classical approach of Section 6.A.

We now consider the general problem of convergence of series of independent random variables. The following will show that the martingale convergence theorem will be powerful. The technique of the proof is other than the result itself will be useful in the development of the theory of applications to sums of dependent random variables. (For example, see Problem 11.)

6.8.4 **Theorem.** *Let $\{X_n, \mathcal{F}_n;\ n = 1, 2, \ldots\}$ be a submartingale and let $Z = \sup_n (X_n - X_{n-1})^+$ (define $X_1 = 0$). If $EZ < \infty$ (for example, if $X_n - X_{n-1} \le M$ a.s.) then $\{X_n\}$ converges to a finite limit a.e. on the set $\{\sup_n X_n < \infty\}$. If $|X_n| \le M$ a.s. and $E|X_1| + E\sum_{n=1}^{\infty} |X_n - X_{n-1}| < \infty$, then X_n converges to a finite limit a.e. on $\{\sup_n X_n < \infty\}$.*

Proof. Fix $M > 0$ and let $T = \inf\{n: X_n > M\}$, and $T = \infty$ if there is no such n. Define $Y_n = X_{n \wedge T} = X_1 + \sum_{k=2}^{n \wedge T} (X_k - X_{k-1})$. Then $\{Y_n\}$ is a submartingale by 6.3.4(ii). This is sometimes called the *martingale stopping theorem*. Since $Y_n = X_n$ if $n \le T$, $Y_n = M$ on $T \le n$. Thus $Y_n^+ \le M + X_{n \wedge T}^+$ is the stopped process stopped at time T.

Now if $n < T$, then $X_n = X_{n \wedge T} = X_1 + (X_2 - X_1) + \cdots + (X_{n \wedge T} - X_{n-1 \wedge T}) \le M + X_1^+ + Z$, and if $n \ge T$ then $X_n = X_{n-1} + (X_T - X_{T-1}) \le M + Z$. Thus $Y_n^+ \le M + 2Z$ in any case, so $\sup_n EY_n^+ \le M + 2EZ < \infty$ by hypothesis. By 6.4.3, Y_n converges a.e. to a finite limit. But if $T = \infty$ then $Y_n = X_n$; hence X_n converges a.e. on $\{\sup_n X_n \le M\}$. Since M is arbitrary, X_n converges a.e. on $\{\sup_n X_n < \infty\}$. The last statement of the theorem is proved by applying the above argument to $-X_n$. ∎

We have seen in 6.1.1 that if $X_1, X_2, \ldots$ are independent random variables with mean 0 and $\sum_{n=1}^{\infty} E(X_n^2) < \infty$, then $\sum_{n=1}^{\infty} X_n$ converges a.e. Kolmogorov's theorem is a partial converse to this result, which we prove after one preliminary

6.8.3 *Theorem.* Let $(Y_1, X_2, \dots)$ be a martingale and $E(Y_n - Y_{n-1})^2 < \infty$ for all n. Then the martingale differences $X_1 = Y_1 - EY_1, \dots, X_n = Y_n - Y_{n-1}, \dots$ are orthogonal.

Proof. Let $j < k$ and $\mathcal{F}_j = \sigma(X_1, \dots, X_j)$. Then

$$E(X_j X_k) = E[E(X_j X_k | \mathcal{F}_{k-1})] = E[X_j E(X_k | \mathcal{F}_{k-1})] = E[X_j E(Y_k - Y_{k-1} | \mathcal{F}_{k-1})]$$

since $X_j = Y_j - Y_{j-1} \in \mathcal{F}_{k-1}$ measurable. But $E(Y_k - Y_{k-1} | \mathcal{F}_j) = Y_j - Y_j = 0$ by the martingale property, and the result follows. □

6.8.4 *Theorem.* Let $Y_1, Y_2, \dots$ be independent random variables with 0 means and assume $\sup_k E Y_k^2 < \infty$. Then (apart from a set of probability 0) the Y_k are uniformly bounded; if $\sum_{k=1}^{\infty} Y_k$ converges a.s., then $\sum_{k=1}^{\infty} E(Y_k^2) < \infty$. (As in Section 6.2, "convergence" of a series means convergence to a finite limit.)

Proof. As in 6.8.3, $Y_n = \sum_{k=1}^{n} Y_k$ forms a martingale. Choose M so that

$$P\left\{ \sup_n |Y_n| \le M \right\} > 0;$$

this is possible since the Y_n converge a.s.

Let $T = \inf\{n : |Y_n| > M\}$ (or ∞ if there is no such n). If $T = \infty$ a.s., then $\{|Y_n| \le M\}$ is a tautology, and (as in 6.8.1) $|E Y_n| \le M$. Now $\sup_n |Y_{n \wedge T}| \le M + |Y_T|$, where $E|Y_T| < \infty$ by hypothesis. It follows that the numbers $E(Y_{n \wedge T})$ are uniformly bounded, so by 6.8.3, $Y_{n \wedge T}$ converges a.s. and in L^2. But by 6.8.3, $E(Y_{n \wedge T})^2 = \sum_{k=1}^{n \wedge T} E(Y_k - Y_{k-1})^2$, with $Y_0 = 0$. Since $Y_{n \wedge T}$ is L^2-convergent, $E(Y_{n \wedge T})^2$ stays bounded as a finite limit, so $\sum_{k=1}^{\infty} E(Y_k - Y_{k-1})^2 < \infty$.

But $X_k = Y_k - Y_{k-1}$, so $\sum_{k=1}^{\infty} E(X_k^2 I_{\{T \ge k\}}) < \infty$, and consequently

$$\sum_{k=1}^{\infty} E(X_k^2 I_{\{T \ge k\}} | X_1, \dots, X_{k-1}) < \infty \quad \text{a.s.}$$

Now we can show that if $X_n \ge 0$ and $\sum_n E(X_n) < \infty$, then $P\{\sum_n X_n < \infty\} = 1$. For $\sum_n X_n < \infty$ a.s. $\Leftrightarrow \sum_n E(X_n | X_1, \dots, X_{n-1}) < \infty$ a.s. Let $Z_n = E(X_n I_{\{|X_n| \le 1\}} | X_1, \dots, X_{n-1})$.

Now $|Z_n| \le 1$, where $A \in \sigma(X_1, \dots, X_{n-1})$; consequently

$$\sum_{n=1}^{\infty} E(Z_n I_A) \le \sum_{n=1}^{\infty} P(A) E(X_n) < \infty \quad \text{a.s.}$$

Pick an ε where $\sum \varepsilon_n$ converges and where $\sup_n |X_n(\omega)| \le M$ (see the beginning of the proof). Then $P(\omega) = \infty > 0$ for all n; thus $\sum_{n=1}^{\infty} \varepsilon_n V_n^2(\omega) < \infty$.

We now have a partial solution to the general problem.

6.8.7 *Theorem.* Let $X_1, X_2, \ldots$ be independent random variables with 0 mean, and suppose $\mathbb{E}[\sup_n |X_n|^2] < \infty$. Then $\sum_{n=1}^{\infty} X_n$ converges a.e. if and only if $\sum_{n=1}^{\infty} \mathbb{E}(X_n^2) < \infty$.

Proof. Apply 6.8.3 and 6.8.6. ∎

We can now complete the solution to the random sums problem (see the discussion after 6.8.1). If $X_n = a_n Y_n$, where the Y_n are independent with $P[Y_n = 1] = P[Y_n = -1] = \frac{1}{2}$, and $\sum_n Y_n$ converges a.e., then $a_n \to 0$, or the X_n are eventually bounded. By 6.8.7, $\sum_n a_n^2 < \infty$.

Incidentally, there is a quick proof of 6.8.6. If $X_n = \sum_{k=1}^{n} Y_k$, where the Y_k are independent with 0 mean and $\sum_{k=1}^{\infty} \mathbb{E}(Y_k^2) < \infty$, then

$$\begin{aligned} \mathbb{E}[|X_n|] &\le \mathbb{E}[|X_n|^2]^{1/2} \\ &= \text{by the Cauchy–Schwarz inequality} \\ &= \left| \sum_{k=1}^{n} \mathbb{E}(Y_k^2) \right|^{1/2} \end{aligned}$$

since the Y_k are uncorrelated with 0 mean, hence orthogonal. Thus $\sup_n \mathbb{E}[|X_n|] < \infty$, but then X_n converges a.e.

The same argument also shows that if $|X_n|$ is uncorrelated and $\sum_{n=1}^{\infty} \mathbb{E}(X_n - X_{n-1})^2 < \infty$, so that the $Z_n = X_n - X_{n-1}$ are orthogonal by 6.8.5, but not necessarily independent, then X_n converges a.e.

We now digress for a proof of von Neumann.

6.8.8 *Theorem.* Let $Y_1, Y_2, \ldots$ be independent, uniformly bounded random variables. Then $\sum_{n=1}^{\infty} Y_n$ converges a.e. if $\sum_{n=1}^{\infty} \mathrm{Var}\, Y_n$ converges and $\sum_{n=1}^{\infty} \mathbb{E}(Y_n)$ converges.

Proof. "If": By 6.8.7, $\sum_{n=1}^{\infty} (Y_n - \mathbb{E}(Y_n))$ converges a.e. since $\sum_{n=1}^{\infty} \mathbb{E}(Y_n - \mathbb{E}(Y_n))^2 = \sum_{n=1}^{\infty} \mathrm{Var}\, Y_n$ converges. Hence $\sum_{n=1}^{\infty} Y_n$ converges a.e.

"Only if" by contradiction. Let $Z_1, Z_2, \ldots$ be independent, where for each n, Z_n has the same distribution as Y_n. Then $\mathbb{E}(Y_n - Z_n) = \mathbb{E}(Y_n) - \mathbb{E}(Z_n) = 0$, $\mathrm{Var}(Y_n - Z_n) = \mathrm{Var}\, Y_n + \mathrm{Var}\, Z_n = 2\,\mathrm{Var}\, Y_n$. Since $\sum_n Y_n$ converges a.e. so does $\sum_n Z_n$. To see this, note that $P[|Y_1 + \cdots + Y_n| \le B] \cdot P[|Z_1 + \cdots + Z_n| \le B] \ge \varepsilon/2$ for all n and all $B > M(\varepsilon)$, hence $P_n \to P_\infty > 0$. It follows that $\sum_n (Y_n - Z_n)$ converges a.e. so by 6.8.7,

$\sum_n$. Var $Y_n < \infty$. Again by 6.3.2, $\sum_n (Y_n - EY_n)$ converges a.s.; hence $\sum_n EY_n$ converges. ∎

Finally, we obtain a general criterion for convergence of series of independent random variables.

6.3.3 *Kolmogorov's Three Series Theorem.* Let $Y_1, Y_2, \ldots$ be independent random variables. Let $A > 0$. Define

$$Y_n' = \begin{cases} Y_n & \text{if } |Y_n| \le A, \\ 0 & \text{if } |Y_n| > A. \end{cases}$$

(a) If $\sum_n Y_n$ converges a.s., then for any $A > 0$, the three series $\sum_n P[|Y_n| \ge A]$, $\sum_n EY_n'$, $\sum_n \mathrm{Var}\, Y_n'$ all converge.

(b) Conversely, if for some $A > 0$, the three series converge, then $\sum_n Y_n$ converges a.s.

Proof. (a) By hypothesis, $Y_n \to 0$ a.s., so eventually $Y_n = Y_n'$. Thus $Y_n \neq Y_n'$ for only finitely many n, that is, $P[\limsup [Y_n \neq Y_n']] = 0$. By the second Borel–Cantelli lemma, $\sum_n P[|Y_n| \ge A] < \infty$. The other two series converge by 6.3.2.

(b) By 6.3.2, $\sum_n Y_n'$ converges a.s. Since $\sum_n P[Y_n \neq Y_n'] < \infty$, we have, almost surely, $Y_n = Y_n'$ eventually, hence $\sum Y_n$ converges a.s. ∎

6.3.4 *Branching Processes.* As a final example we analyze in detail the branching process of 5.3.2(b). Recall that $Z_0 = 1$, and if $Z_n = k$, then $Z_{n+1} = \sum_{i=1}^{k} Y_{ni}$, where $Y_{n1}, \ldots, Y_{nk}$ are independent and $P[Y_{ni} = j] = p_j$, $j = 0, 1, \ldots$. We assume that $m = \sum_{j=0}^{\infty} j p_j$ and $\sigma^2 = \sum_{j=0}^{\infty} (j - m)^2 p_j$ are finite.

This excludes the degenerate case $p_1 = 1$. In this case $Z_n = 1$ for all $n \ge 0$. We also assume that $p_0 + p_1 < 1$. If $p_0 + p_1 = 1$, then $Z_n \ge 1$ for all n. If $p_0 > 0$, then Z_n is eventually 0 since $P[Z_n = 0] \ge p_0^n$ for all n. Finally, if $p_0 = 0$, and if $p_j = 1$, then $Z_n = j^n$.

Case 1. $m < 1$. In this case almost surely Z_n is 0 eventually; thus the family is extinct with probability 1.

For $P[Z_{n+1} = k \mid Z_n = k] = P[Y_{n1} + \cdots + Y_{nk} = k] < 1$, we have $P[Z_{n+1} = Z_n] < 1$. If $Z_n \to 0$, then $Z_n = Z_{n+1}$ infinitely often, and so

$$E\left[\sum_{n=0}^{\infty} Z_n\right] = \sum_{n=0}^{\infty} E[Z_n] < \infty,$$

so $\sum_{n=0}^{\infty} Z_n < \infty$ a.s. But the Z_n are integer-valued, and one with probability 1, Z_n is infinitely often 0.

Case 2. $m = 1$. We show that with probability 1, Z_n is eventually 0, and with probability 1, $Z_n \to \infty$, where ∞ is the unique point in $\mathbb{R}^*$ that is not in $\mathbb{R}$. Note that the latter case has $\sum_{n=0}^{\infty} p_n^{-1} = \infty$.

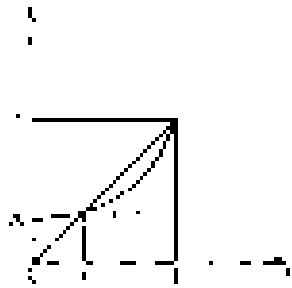

FIGURE 6.4.1.

Let $\phi(x) = \sum_{j=0}^{\infty} a_j x^j$, $0 \le x \le 1$. Consider $\phi(g(x)) - x$. $\phi(x) - x = \phi'(x) - 1$. We have $\phi'(0) = 1 - p_1 = 1 > 0$, $\phi'(1) = 1 - p_1 - 2p_2 - \cdots = \mu - 1 > 0$, and since $p_0 = p_1 = 0$, $\phi'(x) > 0$ is strictly increasing. Since $\phi(x) - x = p_0$ when $x = 0$ and $\phi(1) - 1 = 0$, $g(x) - x$ is strictly decreasing. Thus the minimum occurring somewhere in $[0, 1]$, and $\phi(x) - x$ is strictly increasing to 0 at $x = 1$. It follows that $g(x) < x$ for exactly one $x \in [0, 1)$ when $\mu > 1$ (i.e., $\rho < 1$).

First assume $0 < p_0 < 1$ (thus $p_1 < 1$). By induction $P_n(x) = P(X_n \le x)$ is a nonnegative nondecreasing, and hence convex, function. Since X_n is nonnegative integer-valued, this means that for always every n and every x between zero and the number $P_n(x)$ depending on n is x, $P_n(x) < x$.

Now $P(X_n \text{ eventually constant}) = \sum_{j=0}^{\infty} P(X_n = j)$ for sufficiently large n. If $P(X_n = 1)$ and $P(X_n = 0 \text{ eventually}) > 0$, then $P(X_n > j \text{ for all } n) > 0$ for some j. But by the Markov property, the probability must be $P(X_n = j \mid X_n = j, X_{n-1} = j, \ldots) = p_0^j$. Thus $\alpha = P(X_n = j \text{ for all } n) = p_0^j \alpha$. Since $p_0 < 1$ this implies $P(X_n = 0 \mid X_1 = j) = \alpha = 0$. Thus $\lim_{n \to \infty} q^n = 0$, a contradiction.

Therefore $X_n \to \infty$ or $X_n \to 0$. Since $E(X_n) = \mu^n E(X_0) = \mu^n$ is bounded, the dominated convergence theorem gives $E(\lim_{n \to \infty} X_n) = \lim_{n \to \infty} E(X_n) = 1$. Thus $P(X_n \to \infty) = P(X_n > 0) = 1$. Thus with probability 1, X_n is eventually 0 or with probability 1, $X_n \to \infty$.

If $\mu = 1$, then $\alpha = 0$, hence $X_1 \ge X_2 \ge \cdots$ so that X_n decreases to a limit X. But since the X_n are positive integer-valued

$$\begin{aligned} P(X = \infty) &= P(X_n \quad \text{eventually equal}) \\ &= \sum_{j=1}^{\infty} P(X_1 = j \quad \text{eventually}) \\ &= 0 \quad \text{by the same argument as in the case} \quad 0 < p_0 < 1. \end{aligned}$$

(In the second step we used $P(X_n \quad X \mid X_1 = j) = P(Y_1 + \cdots + Y_j = j) = p_0^j$ where Y_i = number of i's in X_n.)

Case 1: $\nu = 1$. Here, extinction occurs with probability 1. Let $\mu = \sum_{j=0}^{\infty} j p_j$ be the mean of the offspring distribution. Then $\{Z_n/\mu^n\}$ is a nonnegative martingale and hence converges a.s. to a finite limit. The analysis of case 1 shows that Z_n cannot converge to a nonzero constant on a set of positive probability, and hence $Z_n \to 0$ a.s. In other words, with probability 1, the population eventually dies out.

Problems

1. Let $X_1, X_2, \ldots$ be arbitrary random variables with $S_n = X_1 + \cdots + X_n$. What is the relation between the tail σ-fields of the X_n and the tail σ-field of the S_n?
2. What happens in the branching process Exercise 3(b) if instead of $X_1 = 1, X_2$ is an arbitrary nonnegative integer-valued random variable?
3. (Gambler's ruin/fair sandbox model) Let X_n be a gambler's capital after a play of a game of chance. Assume that the gambler has the option of betting or passing at each trial. If he passes at trial $n+1$, then $X_{n+1} = X_n$, and if he bets, then we assume that $|X_{n+1} - X_n| \geq b > 0$. (In this way there is a minimum amount that can be won or lost on a given trial. We do not specify on the gambler's strategy in detail, we simply assume that his strategy coincides with the below rules concerning the distribution of $(X_1, X_2, \ldots)$. It is reasonable to assume that the game is unfavorable or at least "fair" when the rules X_n is a nonnegative supermartingale.

   Let T be the time at the last bet; that is, T equals n such that $|X_{n+1} - X_n| < b$ if $T = \infty$ if there is no such n. Note that T is not a stopping time.

   (a) Show that T is not finite.

   (b) Show that $\sup_n E|X_{n+1} - X_n| = \infty$ in this case, even though the expected value of T is finite.

   (c) If the gambler's strategy is always to bet as long as his capital is at least b, show that $\{X_n\}$ will eventually be constant X_∞, in other words, the gambler will eventually go broke with probability 1.

   (d) If an unbiased coin is tossed independently over and over again, and we own a dollar for each head and owe a dollar for each tail, our accumulated capital is approximately normal. (Exercise 7.) Suppose our strategy is to keep betting until we are ahead by b dollars and then quit. Then our winning is b with probability 1. Why is this not a contradiction?

Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ be increasing σ-fields and $\mathcal{F}_1, \mathcal{F}_2, \ldots$ an increasing sequence of sub-σ-fields of $\mathcal{F}$. Assume that $\mathcal{F}_n \subset \mathcal{F}_k$ measurable for each k and define

$$\mathcal{F}_\infty = \bigvee_{n=1}^{\infty} \mathcal{F}_n = \sigma\left[\bigcup_{n=1}^{\infty} \mathcal{F}_n\right]$$

$\{\mathcal{F}_n\}$ and $\mathcal{F}_n = \sigma\{Y_1, \ldots, Y_n\}$, so that $E[Y_{n+1} \mid \mathcal{F}_n] = E[Y_{n+1} \mid Y_1, \ldots, Y_n]$. Show that $\{X_n, \mathcal{F}_n\}$ is a martingale.

5 (Doeblin's generalization of the Borel–Cantelli lemma) Let $\{\mathcal{F}_n\}$ be an increasing sequence of σ-fields, and let $A_n \in \mathcal{F}_n$, $n = 1, 2, \ldots$. Define

$$q_n = P(A_{n+1} \mid \mathcal{F}_n), \qquad n = 1, 2, \ldots.$$

Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, so that $q_1 = P(A_1)$. Show that $\{\omega : \sum_{n=1}^{\infty} q_n(\omega) = \infty\}$ differs only by a set of probability 0 from $\{\omega : \sum_{n=1}^{\infty} I_{A_n}(\omega) = \infty\}$, that is, $P[\{\omega : \sum_{n=1}^{\infty} I_{A_n}(\omega) = \infty\} \triangle \{\omega : \sum_{n=1}^{\infty} q_n(\omega) = \infty\}] = 0$. Consequently $\sum_{n=1}^{\infty} I_{A_n}$ and $\sum_{n=1}^{\infty} q_n$ have essentially the same convergence set. (Apply Exercise 4 to $X_n = \sum_{k=1}^{n} [I_{A_k} - q_k]$.)

5.9 Applications to Markov Chains

In this section we apply martingale theory to the problem of classifying the states of a (discrete) Markov chain. We begin with a new proof of the basic properties of Markov chains, the reader who is unfamiliar with this subject may consult Ash (1970, Chapter 7). In particular, a state j is said to be an *essential* state if a chain starting at j will return to j with probability 1; otherwise, the state is *inessential*. If C is a set of states such that every state in C can be reached from a finite number of steps from every other state that all states in C are of the same type (essential or inessential).

5.9.1 Lemma on Transient States

5.9.2 *Theorem.* Let $\{Y_n\}$ be the transition records of a Markov chain such that every state in the state space S is reachable from every other state (irreducible chains are called *recurrent* chains). Choose a fixed state, and label it 0 for convenience. The states are transient if there is a nonconstant bounded $g : S \to R$ such that $\sum_{k} p_{jk} g(k) = g(j)$ for all $j \neq 0$.

Proof Suppose such a g exists. By adding a constant to g we may assume that $g(j) \geq 0$. Assume the initial state is $j \neq 0$, and let $\{X_n\}$ be the corresponding sequence of random variables. Let T be the time at which 0 is reached, and let $Y_n = X_{n \wedge T}$. Then $\{Y_n\}$ can be shown to be a Markov chain with the same initial distribution and transition matrix as $\{X_n\}$, except that 0 is now an absorbing state. In other words, the transition matrix for $\{Y_n\}$ is

$$\begin{aligned} p_{ij}' &= p_{ij} \quad \text{if } i \neq 0 \text{ and } j \neq 0, \\ p_{0j}' &= \end{aligned}$$

Thus $\sum_{j\in S} p_{ij} f(j) = f(i)$ for $i \in C$. In matrix form, $\prod_C f = 1$. By induction $\prod^n f = f$ for all n. Hence

$$\sum_{j\in S} p_{ij}^{(n)} f(j) = f(i),$$

where $p_{ij}^{(n)} = P(X_n = j \mid X_0 = i)$. This implies that $\pi_i f(X_n) \mid X_0 = i] = f(i)$.

If the states of the subset C are recurrent, then C will be visited with probability 1 since $P_C = 1$ on C. By the dominated convergence theorem $E[f(X_n) \mid X_0 = i] \to f(i)$. We conclude that $f(i) = f(j)$ for all i contradicting the hypothesis that f is nonconstant.

Conversely, if the states are transient, we define $f : S \to R$ as follows. If $i \notin C$, let $f(i) = P_{iC}$, the probability that starting from i, X will eventually be reached; take $f(i) = 1$. Now if we take one step to reach C from $i \notin C$, we may either go directly to one step, or go to some $j \notin C$ and then reach C in some time after the first step. It follows that

$$f(i) = \sum_{j\in C} p_{ij} f(j) + \sum_{j\notin C} p_{ij} f(j), \qquad i \notin C.$$

(This may be formalized using the Markov property.)

Since f is clearly bounded, and $f = 1$ on C, it is easy to see $f \neq C$ (otherwise f would be a constant). Thus f is nonconstant.

Martingale theory is used in defining the following sufficient condition for recurrence.

6.9.2 *Theorem.* Let $\{p_{ij}\}$ be the transition matrix of an irreducible Markov chain whose state space S is the set of nonnegative integers. If there is a function $f : S \to R$ such that $f(i) \to \infty$ as $i \to \infty$, and $\sum_j p_{ij} f(j) \leq f(i)$ for all $i \notin C$, then the chain is recurrent.

Proof. As $f(i) \to \infty$ as $i \to \infty$, f is bounded below, so without loss of generality we may assume $f \geq 0$. Let the initial state $X_0 = i$, and form the process $\{Z_n\}$ as in 6.9.1. Since $\sum_{j\in S} p_{ij} f(j) \leq f(i)$ for all i, which implies that $E[Z_n] \leq$ is a nonnegative supermartingale, and hence converges a.e. to a finite limit. For

$$E[Z_n \mid Z_1 = i_1, \ldots, Z_{n-1} = i_{n-1}] = E[f(X_n) \mid X_{n-1} = i_{n-1}]$$

$$= \text{by the Markov property}$$

$$= \sum_j p_{i_{n-1} j} f(j)$$

$$\leq f(i_{n-1}).$$

Note also that

$$\sum_{k=0}^{\infty} k\,\mathbb{P}(X_n = k) = \sum_{k=0}^{\infty} k\,p_k^{(n)} \to \sum_{k=0}^{\infty} k\,\pi_k \quad \text{as } n \to \infty,$$

because $|p_k^{(n)} - \pi_k| \to 0$.

Assume the states transient. Then $p_{ij}(n) \to 0$ for some $i, j \in \mathbb{N}$. Choose such i as the initial state. Now $X_n \to \infty$ as $n \to \infty$ since a finite set of transient states cannot be visited infinitely often (see Ash (1972, p. 207)). This means that with probability one, $X_n > 0$ for $n > N$. But $\mathbb{P}(X_n > 0) = 1 - p_{i0}^{(n)} \to 1$ and hence $\mathbb{P}(X_n > 0) \to 1 > 0$; this implies that $\mathbb{P}(X_n) > 0 \to 0$, a contradiction. □

The proof of 6.9.2 shows that if $\mathbf{P} = (p_{ij})$ is the transition matrix of a Markov chain (X_n), f is a real-valued function on the state space and

$$\sum_i p_{ij} f(i) = f(j) \qquad \text{for all } j$$

where the series is assumed to converge absolutely, then with i fixed in such a way that $\mathbb{P}(X_0 = i) > 0$, the sequence $(f(X_n))$ is a supermartingale. Similarly replacement of "$\geq$" by "$=$" in this equation yields a martingale, and replacement by "$\leq$" yields a submartingale.

We now apply 6.9.1 and 6.9.2 to a queueing problem.

6.9.3 **Example.** Assume that customers arrive to be served at discrete times $n = 0, 1, \ldots$, and at most one customer can be served at a given time. Suppose there are X_n customers waiting for completion of service at time n, and in the interval $[n, n+1)$, Y_n new customers arrive, where $\mathbb{P}(Y_n = k) = p_k$, $k \geq 0$. The number of customers before completion of service at time $n+1$ is

$$X_{n+1} = (X_n - 1)^+ + Y_n.$$

The queueing process may be represented as a Markov chain whose state space is the set of nonnegative integers and whose transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & \cdots & \end{array} \\ \left|\begin{array}{cccccc} p_0 & p_1 & p_2 & p_3 & \cdots & \\ p_0 & p_1 & p_2 & p_3 & \cdots & \\ 0 & p_0 & p_1 & p_2 & p_3 & \cdots \\ 0 & 0 & p_0 & p_1 & p_2 & \cdots \\ & & & \vdots & & \end{array}\right| \end{array}.$$

We assume that $p_0 > 0$ and $p_0 + p_1 < 1$, so that the chain is irreducible. We shall prove the existence of the chain in Example 6.10.2.

The equations $\sum_{n=0}^{\infty} p_n P(i) = P(i), i \geq 0$, are

$$\alpha_0(i) = 1: \quad p_0 P(0) + \alpha_0 P(i-1) + \cdots + p_{i-1} P(1) + (1 - p_1 + \cdots - P(i) \tag{15}$$

Take $\alpha = \sum_{i=0}^{\infty} \alpha_i P(i) = \sum_{i=0}^{\infty} \alpha_i$. Then we have the following two possibilities: $\alpha > 1$ recurrent if $\alpha < 1$.

Now assume $\alpha = 1$; if $P(i) = \alpha^i$ then we have

$$p_0 P(i-1) - p_1 P(i) + p_2 P(i-2) - \cdots - p_{i-1} P(1) + (1 - \cdots - P(i)$$

or

$$\sum_{n=0}^{i} \alpha_n P(i-n) = \alpha_i$$

But this is the substitution of $\alpha_i = P(i)$ into case 2 of (6.5.10). Thus $P(i)$ is bounded and non-constant $\alpha = \sum p_i = 1$, the state is transient.

Now assume $\alpha = 1$, and let $P(i) = P$. Then $P = \alpha$.

$$\sum_{i=0}^{\infty} p_i P(i) \quad p_0 P(i-1) + p_1 P(i-2) + \cdots + p_{i-1} P(1) + \cdots$$

$$- \sum_{n=0}^{i-1} p_n P(i-n-1)$$

$$- \sum_{n=0}^{i-1} (\alpha_n - \alpha_{n-1}) P(i-n-1)$$

$$- \sum_{n=0}^{i-1} \alpha_n P(i-n-1) + \alpha_{i-1} P(1) - \alpha_i P(0)$$

By (6.9.2), this is also zero.

Thus $\alpha = 1$ implies recurrence; μ_i is the average length of time required to return to i when the initial state is i. State i is said to be recurrent null if $\sum p_i = \infty$, otherwise positive recurrent. Thus we have shown that the state is recurrent null if $\alpha = 1$, recurrent positive if $\alpha < 1$ (see Karlin [Wo, pp. 216]).

Problems

1 Consider the Markov chain in Section 6.2, Problem 2, and assume $\alpha = 1$. If $\alpha = 1$ as in the above, show that $X_n = \cdots$ for $n \to \infty$, hence the state is transient.

6.10 Martingales

Many classical results on strong laws of large numbers, stationarity and related topics may be found in Feller (1950) and Doob (1953). Another useful reference is Breiman (1968).

Martingales and Markov processes are important in the area of probabilistic potential theory. For an introduction to this subject, see Meyer (1966) and Kemeny, Snell, and Knapp (1976).

Miscellaneous works are by Shiryaev (1996) and Williams (1991).

CHAPTER

7

THE CENTRAL LIMIT THEOREM

7.1 Introduction

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with common mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ be the sum of the first n of these variables. The law of large numbers tells us that S_n/n converges to μ in probability. This gives us information about the distribution of S_n for large n; in other words, it tells us what is happening to the center of the distribution of S_n when n is large. The central limit theorem gives us information about the distribution of S_n for large n. The purpose of this chapter is to develop results leading to one of the most important theorems concerning the approximate distribution of S_n for large n. For example, if the X_i are integer-valued with finite mean μ and finite variance σ^2, then for large n, $P(S_n = k)$ is approximately $\frac{1}{\sigma\sqrt{n}}$ times the ordinate of the normal distribution with mean $n\mu$ and variance $n\sigma^2$.

There are two main techniques that will be used. First is the theory of weak convergence. If $\mu_1, \mu_2, \ldots$ are finite measures on $\mathcal{B}(R)$, weak convergence of μ_n to μ means that $\int_R f\,d\mu_n \to \int_R f\,d\mu$ for every bounded continuous f on R. If the corresponding distribution functions are $F, F_1, F_2, \ldots$, the equivalent condition is that $F_n(x) \to F(x)$ for all x in R that are continuity points of F. (See 2.8 for a discussion of weak convergence in probabilistic terms.) Note that $-\infty$ and $+\infty$ are, by definition, continuity points of F. We shall denote weak convergence by $\mu_n \Rightarrow \mu$ or $F_n \Rightarrow F$. Also, $\mathcal{C}(F)$ is a Borel subset of R, the union $D(F)$ and $\mathcal{C}(F)$ will be complementary.

Now suppose that μ_n is the probability measure induced by a random variable X_n, $n = 0, 1, \ldots$. If $\mu_n \Rightarrow \mu_0$, $Y_n = X_n$ a.s. $[\mu_n]$, $n \geq 0$, and the sequence $\{X_n\}$ converges in distribution to X_0 and with $X_n \to Y_n$. Since $\mu_{Y_n} = \mu_n$, it follows that $X_n \to X_0$ a.s. $[P]$ and $E[f(X_n)] \to E[f(X_0)]$ for all bounded continuous $f: S \to R$.

This in turn implies that if $X_n \xrightarrow{d} X$ and g is a continuous function from $\mathbb{R}$ to $\mathbb{R}$, then $g(X_n) \xrightarrow{d} g(X)$. In particular, if $X_n \xrightarrow{d} X$, then $X_n - c \xrightarrow{d} X - c$ for each real number c.

Notice that convergence in distribution is determined completely by the distribution functions or equivalently by the induced probability measures of the random variables. In particular, the random variables need not be defined on the same probability space. Now step functions are discontinuous at the jumps, so the distribution function of a random variable always has the value 0 at $-\infty$ and the value 1 at ∞, we have $X_n \xrightarrow{d} X$ if $F_n(x) \to F(x)$ as $n \to \infty$ for each continuity point x of F on $\mathbb{R}$.

Now by Theorem 9.3.1 $\mu_n \xrightarrow{c} \mu$ iff $\mu_n(A) \to \mu(A)$ for each Borel set A whose boundary ∂A has μ measure 0. Thus $X_n \xrightarrow{d} X$ if $P[X_n \in A] \to P[X \in A]$ for all Borel sets A such that $P[X \in \partial A] = 0$. This result justifies the intuitive idea of convergence in distribution, that it says roughly that if $X_n \xrightarrow{d} X$, then X_n and X have approximately the same distribution for large n. Of course it is much more beautiful to realize that $P[X_n \in A] \to P[X \in A]$ but not for all sets A, but certainly this is not true. For example, if X_n is uniformly distributed between 0 and $\frac{1}{n}$ [so X_n has density $f_n(x) = n$ if $0 < x < \frac{1}{n}$ and $f_n(x) = 0$ elsewhere], then the large n, X_n approximates a random variable X that is identically 0. But $P[X_n = 0] = 0$ for all n, and $P[X = 0] = 1$.

The second technique involves the use of characteristic functions, which we now define.

7.1.1 *Definition.* Let μ be a finite measure on $\mathscr{B}(\mathbb{R})$. The *characteristic function* of μ is the mapping from $\mathbb{R}$ to $\mathbb{C}$ given by

$$\hat{\mu}(t) = \int_{\mathbb{R}} e^{itx} \mu(dx), \qquad t \in \mathbb{R}$$

where i is the complex number $\sqrt{-1}$. If F is a distribution function corresponding to μ, we shall also write $\hat{F}(t) = \int_{\mathbb{R}} e^{itx} dF(x)$ and call $\hat{F}$ the characteristic function of F and of X if X is a random variable with distribution function F.

Characteristic functions are uniquely appropriate in the study of sums of independent random variables, because of the following result.

7.1.2 *Theorem.* Let $X_1, X_2, \ldots, X_n$ be independent random variables, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then the characteristic function of S_n is the product of the characteristic functions of the X_k.

Hence

$$E(e^{it\bar{X}_n}) = E\left[\prod_{j=1}^{n} e^{itX_j/n}\right]$$
$$= \prod_{j=1}^{n} E(e^{itX_j/n}) \qquad \text{by independence,}$$

Theorem 7.1.3 allows us to compute the characteristic function of $\bar{X}_n$ knowing only the distribution of the individual X_j's. In fact, since the characteristic function is known, the distribution function is determined.

9.4.1 *Inversion formula.* If ϕ is the characteristic function of the bounded distribution function F, and $F(a,b] = F(b) - F(a)$, then

$$F(a,b] = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi(t)\,dt$$

for all points $a < b$ at which F is continuous. If the integral on the right is absolutely convergent, then the function f given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\phi(t)\,dt$$

is a density for F, that is, f is nonnegative and $F(x) = \int_{-\infty}^{x} f(t)\,dt$ for all x; furthermore, $F'(x) = f(x)$ everywhere. Thus in this case f and ϕ are Fourier transform pairs:

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx,$$
$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx}\phi(t)\,dt.$$

If we are trying to compute the distribution of a sum S_n of independent random variables X_j, then Theorem 9.4.2 will be useful only if we can compute the distribution function from the characteristic function. In fact this is the case:

Example. Suppose $g(x) = 1/\pi(1+x^2)$, then $\hat{g}(t) = \int_{-\infty}^{\infty} e^{itx} g(x)\,dx$ is $e^{-|t|}$, so

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-|t|} e^{-itx}\,dt.$$

Thus

$$F(a,b] = \int_{a}^{b} g(x)\,dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} e^{-|t|}\,dt.$$

and since

$$\int_a^b e^{-ixy}\,dx = \frac{e^{-iby} - e^{-iay}}{-iy},$$

the validity of the inversion formula. For a formal proof, let

$$I_{\varepsilon,N} = \frac{1}{2\pi}\int_{-N}^{N} \frac{e^{-iay} - e^{-iby}}{iy} \hat{f}(y)\,e^{-\varepsilon|y|}\,dy, \qquad \varepsilon > 0.$$

Then

$$I_{\varepsilon,N} = \frac{1}{2\pi}\int_{-N}^{N} e^{-iay}\,e^{-\varepsilon|y|} \left| \int_{-\infty}^{\infty} e^{iyx} f(x)\,dx \right| dy.$$

Now

$$\left| \int_{-N}^{N} e^{-iay}\,e^{-\varepsilon|y|} \int_{-\infty}^{\infty} e^{iyx} f(x)\,dx \right| \le \int_{-N}^{N} e^{-\varepsilon|y|} \left| \int_{-\infty}^{\infty} e^{iyx} f(x)\,dx \right| dy \le A(\varepsilon) < \infty,$$

and

$$\int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} e^{iyx} f(x)\,dx \right| dy = 2\pi \|\hat{f}\|_{L^1(\mathbb{R})} < \infty.$$

By Fubini's theorem, the order of integration may be interchanged to obtain

$$I_{\varepsilon,N} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \int_{-N}^{N} e^{-iay}\,e^{-\varepsilon|y|} \left(\frac{e^{iyx} - e^{iyb}}{iy} \right) dy\,f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx\,F_{\varepsilon,N}(x),$$

where

$$F_{\varepsilon,N}(x) = \frac{1}{2\pi}\int_{-N}^{N} \left(\frac{\sin y(x-a)}{y} - \frac{\sin y(x-b)}{y} \right) e^{-\varepsilon|y|}\,dy.$$

Note that

$$\int_{-N}^{N} \left(\frac{\sin y(x-a)}{y} - \frac{\sin y(x-b)}{y} \right) dy = 0$$

since the integrand is an odd function. Let $u = y(x-a)$ and $v = y(x-b)$. It follows that

$$F_{\varepsilon,N}(x) = \frac{1}{2\pi}\int_{-N(x-a)}^{N(x-a)} \frac{\sin u}{u}\,du - \frac{1}{2\pi}\int_{-N(x-b)}^{N(x-b)} \frac{\sin v}{v}\,dv.$$

Now

$$\int_0^{\infty} \frac{\sin u}{u}\,du = \frac{\pi}{2} \quad \text{and} \quad \int_{-\infty}^{0} \frac{\sin u}{u}\,du = -\frac{\pi}{2},$$

and since this integral is continuous in t and x, it is bounded uniformly in x and n. Thus for some $M < \infty$, $|f_n(x)| \le M$ for all x and n, but if x is fixed as $n \to \infty$, $f_n(x) \to f(x)$, where

$$f(x) = \begin{cases} 0 & \text{if } x \text{ is } \\ & \text{if } \\ & \text{if } \end{cases}$$

By the dominated convergence theorem,

$$\begin{aligned} \lim_{n\to\infty} I_n &= \int_{-\infty}^{\infty} f(x)\,dF(x) \\ &= \tfrac{1}{2}[F(b+) - F(a+)] + \tfrac{1}{2}[F(b-) - F(a-)] \\ &= \frac{F(b) + F(b-)}{2} - \frac{F(a) + F(a-)}{2} = F(b) - F(a) \end{aligned}$$

if F is continuous at a and b. This proves the formula.

Now assume h is integrable on $(-\infty, \infty)$. Let

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ixt} h(t)\,dt \qquad \text{for } -\infty < x < \infty.$$

Since h is integrable, f is well-defined, continuous, f is continuous by the dominated convergence theorem. Now by Fubini's theorem,

$$\begin{aligned} \int_a^b f(x)\,dx &= \frac{1}{2\pi} \int_{-\infty}^{\infty} h(t) \left[\int_a^b e^{-ixt}\,dx \right] dt \\ &= \lim_{n\to\infty} \frac{1}{2\pi} \int_{-n}^{n} h(t) \left[\int_a^b e^{-ixt}\,dx \right] dt \\ &= \lim_{n\to\infty} \frac{1}{2\pi} \int_{-n}^{n} \frac{e^{-ibt} - e^{-iat}}{-it} h(t)\,dt \\ &= F(b) - F(a) \end{aligned}$$

by the inversion formula if a and b are continuity points of F. Thus $F(b) - F(a) = \int_a^b f(x)\,dx$ at continuity points of F.

Define every point as a limit from above or continuity points since F is monotone and thus has only countably many discontinuities. Since the integral is continuous, F is the integral of its derivative. It follows that $F(b) - F(a) = \int_a^b f(x)\,dx$ for all a and b.

Since F is continuous, $\{\text{the set on which } F \text{ is differentiable}\}$ is a polish set and its complement is μ. Since F is increasing, f is everywhere continuous. Thus f is a density for μ. $\square$

7.1.4 *Corollary.* *For F and G probability measures on $\mathbb{R}$ (or more generally, finite measures on $\mathcal{B}(\mathbb{R})$), if $\int_{\mathbb{R}} e^{tx}\,dF(x) = \int_{\mathbb{R}} e^{tx}\,dG(x)$ for all $t \in \mathbb{R}$, then $F = G$.*

Proof. By the inversion formula, $F(b) - F(a)$ is determined for all continuity points. But as in the proof of 7.1.3, every point is a limit from above of continuity points, and it follows that F is determined everywhere. $\square$

Various procedures involving Fourier or Laplace transforms may be used in actually computing the distribution of a random variable from its characteristic function (see Feller, 1970, Chapter 5).

Characteristic functions have the following basic properties.

7.1.5 *Theorem.* *Let φ be the characteristic function of the bounded distribution function F. Then*

(a) $|\varphi(t)| \le 1 = \varphi(0) = \varphi(-t)$ for all t;

(b) φ is continuous on $\mathbb{R}$;

(c) $\varphi(-t) = \overline{\varphi(t)}$, the complex conjugate of $\varphi(t)$;

(d) $\varphi(at+b) = e^{ibt}\varphi(at)$ for all a, b iff F is symmetric; that is, $\int_A dF(x) = \int_A dF(-x)$ for all Borel sets $A \subset \mathbb{R}$, where $-A = \{-x : x \in A\}$.

(e) If $\int_{\mathbb{R}} |x|^k\,dF(x) < \infty$ for some positive integer k, then the kth derivative of φ exists and is continuous on $\mathbb{R}$, and

$$\varphi^{(k)}(t) = \int_{\mathbb{R}} (ix)^k e^{itx}\,dF(x).$$

Proof. Part (a) is clear since $e^{it} = 1$, and (b) follows from the dominated convergence theorem (see Problem 7.1.5). Part (c) follows from the fact that the conjugate of e^{it} is e^{-it}. To prove (d), assume A to be real-valued and, for now, let F be the distribution function of the random variable X. Now $E(e^{it(-X)}) = E(\cos t(-X)) + iE(\sin t(-X)) = E(\cos tX) - iE(\sin tX)$, which is real, so that $-X$ has characteristic function φ by (c). But $-X$ has the same distribution, hence $P[X \in A] = P[-X \in A]$, and F is therefore symmetric. In general, we must consider F to be a positive constant times a probability P with $P(\mathbb{R}) = 1$. (If $F(\infty) = 0$, then $F(x) = 0$ for all x, and the result is trivial.) The above argument shows that $\int_B \cos tx\,dF(x) = \int_B \cos tx\,dF(-x)$, and hence F is symmetric.

Conversely, assume F symmetric, and let $g: \mathbb{R} \to \mathbb{R}$ be a bounded Borel measurable function. If g is odd, $g(-x) = -g(x)$ for all x, then

$\int_{\mathbb{R}} g(x)\,dF(x) = 0$ for all t to prove that g is the zero function. In particular, $\int_{\mathbb{R}} \sin(tx)\,dF(x) = 0$ and h is real-valued.

Finally, we prove (a). Since $\int_{\mathbb{R}} \cos(tx)\,dF(x) = \frac{1}{2}$ and $\int_{\mathbb{R}} \sin(tx)\,dF(x) = 0$, we may differentiate $\psi(t) = \int_{\mathbb{R}} e^{itx}\,dF(x)$ times under the integral sign (see 16.6, Problem 3 and 2.4, Problem 10), and the result follows. ∎

Now suppose that h is a given function from $\mathbb{R}$ to $\mathbb{C}$, and we wish to determine whether or not h is the characteristic function of some bounded distribution function F. There are some practical approaches to this difficult problem. For example, if h fails to satisfy $|h(t)| \le h(0)$ for some t, h cannot be a characteristic function. Also, suppose for contradiction that Lévy's inversion formula (see 16.6, Problem 3) holds for all a and b, so that

$$F(b) - F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} h(t)\,dt, \qquad a < b.$$

If h is Lebesgue integrable, then

$$\lim_{a \to -\infty} F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ita} \frac{h(t)}{it}\,dt, \qquad a \in \mathbb{R}.$$

(This is a standard Fourier transform result; see Rudin, 1966, p. 156.) Thus if h is everywhere nonnegative, then h is the characteristic function of a measure μ with density f, that is,

$$\mu(B) = \int_B f(x)\,dx, \qquad B \in \mathcal{B}(\mathbb{R}).$$

There is a general criterion for deciding whether or not h is a characteristic function. The result is of considerable importance in the development of certain deeper properties of stochastic processes. However, it is inconvenient when applied to explicit examples.

7.1.5 Bochner's Theorem. The function $h: \mathbb{R} \to \mathbb{C}$, $h(0) = 1$, is a characteristic function iff h is continuous at the origin and is *nonnegative definite*, in other words for all $t_1, \ldots, t_n \in \mathbb{R}$, $n = 1, 2, \ldots$, and all complex numbers $z_1, \ldots, z_n$,

$$\sum_{j=1}^{n} \sum_{k=1}^{n} z_j \bar{z}_k h(t_j - t_k) \ge 0$$

(a real and nonnegative

Proof. Suppose φ is a characteristic function. Then λ is nonnegative definite by 7.1.5(ii). To prove nonnegative definiteness, write

$$\sum_{r,s=1}^{n} a_r \bar{a}_s \lambda(t_r - t_s) = \int_{-\infty}^{\infty} \left| \sum_{r=1}^{n} a_r e^{it_r x} \right|^2 dF(x) \quad \left(= \sum_{r,s=1}^{n} a_r \bar{a}_s \int_{-\infty}^{\infty} e^{i(t_r - t_s)x} dF(x) \right)$$

$$= \sum_{r,s=1}^{n} a_r \bar{a}_s \varphi(t_r - t_s) \geq 0.$$

The converse is considerably more difficult, and since the result will not be used in the sequel, the argument will be omitted. For a complete proof see Lukacs (1960) or Ash and Gardner (1975).

We shall need the result that if X is normally distributed with mean μ and variance σ^2, the characteristic function of X is

$$\varphi(t) = \exp(i\mu t) \exp(-\tfrac{1}{2}\sigma^2 t^2).$$

For the computation see Ash (1970), p. 262.

Another basic result for relating the various types of convergence is Slutsky's theorem.

7.1.7 Theorem (Slutsky) (i) If X_n converges to X in probability, then X_n converges to X in distribution. (ii) A partial converse: If X_n converges in distribution to a constant c, then X_n converges in probability to c.

Proof. (i) Let F_n be the distribution function of X_n, and F the distribution function of X. Then

$$F_n(x) = P[X_n \leq x] = P[X_n \leq x, X \leq x + \varepsilon] + P[X_n \leq x, X > x + \varepsilon]$$

$$\leq P[X \leq x + \varepsilon] + P[|X_n - X| > \varepsilon]$$

$$= F(x + \varepsilon) + P[|X_n - X| > \varepsilon], \quad \text{and}$$

$$F(x - \varepsilon) = P[X \leq x - \varepsilon] = P[X \leq x - \varepsilon, X_n \leq x] + P[X \leq x - \varepsilon, X_n > x]$$

$$\leq P[X_n \leq x] + P[|X_n - X| > \varepsilon]$$

$$= F_n(x) + P[|X_n - X| > \varepsilon]. \quad \text{Thus}$$

$$F(x - \varepsilon) - P[|X_n - X| > \varepsilon] \leq F_n(x) \leq F(x + \varepsilon) + P[|X_n - X| > \varepsilon]$$

$$F(x - \varepsilon)$$

Since $X_n \xrightarrow{P} X$, we have $P[|X_n - X| > \varepsilon] \to 0$ as $n \to \infty$. If F is continuous at x, then $F(x - \varepsilon) \to F(x)$ and $F(x + \varepsilon) - P[|X_n - X| > \varepsilon] \to F(x)$. Thus $F_n(x)$ is

bound between two quantities that can be made arbitrarily close to $F(x)$, so $F_{X_n}(x) \to F(x)$.

$$\begin{aligned} F_{X_n}(x) &= P(X_n \le x) = P(Y_n \le x, X_n \le Y_n + \varepsilon) + P(X_n \le x, X_n > Y_n + \varepsilon) \\ &\le F_{Y_n}(x + \varepsilon) + P(|X_n - Y_n| > \varepsilon) \end{aligned}$$

Now $P(Y_n \le x - \varepsilon) = P(Y_n \le x - \varepsilon, X_n \le x) + P(Y_n \le x - \varepsilon, X_n > x)$, so

$$\begin{aligned} P(Y_n \le x - \varepsilon, X_n \le x) &\le P(X_n \le x) \\ P(Y_n \le x - \varepsilon) - P(Y_n \le x - \varepsilon, X_n > x) &\le F_{X_n}(x) \\ F_{Y_n}(x - \varepsilon) - P(|X_n - Y_n| > \varepsilon) &\le F_{X_n}(x) \end{aligned}$$

Because $Y_n \to X$, $F_{Y_n}(x)$ converges to the distribution function of the random variable X, that is,

$$u(x) = \begin{cases} 1 & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Since $F_{Y_n}(x + \varepsilon) \to 1$ and $F_{Y_n}(x - \varepsilon) \to 0$ as $n \to \infty$, and, therefore, $P(|X_n - Y_n| > \varepsilon) \to 0$ as $n \to \infty$. ∎

Problems

1. Give an example of a sequence of random variables X_n such that X_n converges in distribution to X, but X_n does not converge in probability to X.
2. The following application of Theorem 7.1.4 is typical in computations involving characteristic functions. Let f and g be nonnegative Borel measurable functions such that $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} g(x)\,dx = m$, and $\int_{-\infty}^{\infty} f(x)e^{ux}\,dx < \infty$, $\int_{-\infty}^{\infty} g(x)e^{ux}\,dx < \infty$ for all u in some real interval I. If

$$\int_{-\infty}^{\infty} f(x)e^{ux}\,dx = \int_{-\infty}^{\infty} g(x)e^{ux}\,dx \quad \text{for all} \quad u \in I$$

show that $f = g$ except on a set of Lebesgue measure zero.

3. If h_1 and h_2 are characteristic functions, show that $\alpha_1 h_1 + \alpha_2 h_2$ and $h_1 h_2$ are also characteristic functions if $\alpha_1 \ge 0$, $\alpha_2 \ge 0$, and $\alpha_1 + \alpha_2 = 1$. Is $h_1 - h_2$ a characteristic function?
4. Let h be the characteristic function of the random variable X.
   (a) If $h(u) = 1$ for some $u \ne 0$, show that X has a lattice distribution, that is, with probability 1, X belongs to the set $\{a + nb : n = 0, \pm 1, \pm 2, \ldots\}$ for appropriate a and $b > 0$. (Hint: $e^{ix} = 1 \Leftrightarrow x = 2\pi k$ for some integer k.) Conversely, if X has a lattice distribution, then $|h(u)| = 1$ for some $u \ne 0$.
   (b) If $|h(u)| = 1$ for two distinct points u and v whose ratio is irrational, show that X is degenerate, that is, has a certain

5. Let X be a random variable with $E(|X|^n) < \infty$ for some positive integer n. If $\phi(t)$ is the characteristic function of X, show that $\phi^{(k)}(0) = i^k E(X^k)$, $k = 1, 2, \ldots, n$, and

$$\phi(t) = \sum_{k=0}^{n} \frac{E(X^k)}{k!}(it)^k + o(|t|^n),$$

$$\text{cov}(e^{tX}, e^{tY}) = \sum_{k=1}^{n} \frac{\text{cov}(X^k, Y^k)}{k!}(it)^k + \frac{|t|^n \varepsilon_1(t)}{n!},$$

Solution: If X is a random variable with $E(|X|^n) < \infty$ for some $n > 0$, show that the characteristic function of X is given by

$$\phi(t) = \sum_{k=0}^{n} \frac{E(X^k)}{k!}(it)^k$$

within the interval of convergence of the series. This is the moment generating property of characteristic functions.

(b) Give an example of a random variable X with $E(|X|^n) < \infty$ for all $n > 0$, such that the series

$$\sum_{k=0}^{\infty} \frac{E(X^k)}{k!}(it)^k$$

converges only at $t = 0$.

7. (a) Let ϕ be the characteristic function of the bounded distribution function F. That is,

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x), \qquad -\infty < t < \infty.$$

Show that

$$\frac{1}{2T}\int_{-T}^{T} |\phi(t)|^2\, dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\sin T(x-y)}{T(x-y)}\, dF(x)\, dF(y).$$

[Hint: $|\phi(t)|^2 = \phi(t)\overline{\phi(t)}$, where $\overline{\phi(t)} = \int_{-\infty}^{\infty} e^{-ity}\, dF(y)$. Interchange the order of integration and integrate with respect to t.]

(b) If $\phi(t)$ exists and is finite at the origin, show that $\int_{-\infty}^{\infty} x^2\, dF(x) < \infty$. [Use l'Hôpital's rule and Fatou's lemma.]

(c) If $\phi''(0)$ exists and is finite, show that $\int_{-\infty}^{\infty} x^2\, dF(x) < \infty$ for $n = 1, 2, \ldots$.

[It is probably easiest to use part (a) and induction arguments rather than a classical proof.]

5 If X is a random variable, recall $\Psi(t) = E(e^{tX})$, a complex number for $t = -ip$ we obtain the moment generating function of X. If X is analytic at the origin above and $E(|X|^n) < \infty$ for all $n > 0$, and

$$M(t) = \sum_{n=0}^{\infty} \frac{1}{n!} E(X^n) t^n$$

within the circle of convergence of the series. In particular, $M^{(n)}(0) = E(X^n)$.

7.2 The Fundamental Weak Convergence Theorem

The basic connection between weak convergence and characteristic functions is essentially due to Lévy. Let $\{F_n\}$ be a bounded sequence of distribution functions on $\mathbb{R}$ ("bounded" means that for some positive M, $1 - F_n(x) + F_n(-x) \leq M/x$ for all $x > 0$ and all n). Let $\{\phi_n\}$ be the corresponding sequence of characteristic functions. If F is a bounded distribution function with characteristic function ϕ, then weak convergence of F_n to F is equivalent to pointwise convergence of ϕ_n to ϕ. In the course of developing this result, we must consider the following question. If $\{F_n\}$ is a bounded sequence of distribution functions, must there exist a weakly convergent subsequence? Now any bounded sequence of real numbers has a convergent subsequence, so one might conjecture that any bounded sequence of distribution functions has a weakly convergent subsequence. In fact this is not true, but the following result comes close in a sense.

7.2.1 *Helly's Theorem.* Let $F_1, F_2, \ldots$ be distribution functions on $\mathbb{R}$. Assume that $F_n(-n) \leq \varepsilon$ for all n, and $F_n(n) \geq 1 - \varepsilon$ for some ε. Then there is a distribution function F and a subsequence $\{F_{n_k}\}$ such that $F_{n_k}(x) \to F(x)$ for each $x \in \mathbb{R}$ at which F is continuous.

Proof. Let $D = \{r_1, r_2, \ldots\}$ be a countable dense subset of $\mathbb{R}$. Since the sequence $\{F_n(r_1)\}$ is bounded, we can extract a subsequence $\{F_{1,n}\}$ of $\{F_n\}$ with $F_{1,n}(r_1)$ converging to a limit L_1. Since $\{F_{1,n}(r_2)\}$ is bounded, there is a subsequence $\{F_{2,n}\}$ of $\{F_{1,n}\}$ such that $F_{2,n}(r_2)$ approaches a limit L_2. Continuing inductively, we find subsequences $\{F_{k,n}\}$ of $\{F_{k-1,n}\}$ with $F_{k,n}(r_k) \to L_k$, $n = 1, 2, \ldots$, for all k, and L_k is bounded by M/r_k.

Define $F_D : D \to \mathbb{R}$ by $F_D(r_k) = L_k$, $k = 1, 2, \ldots$, and let $F_n = F_{n,n}$, $n = 1, 2, \ldots$, the "diagonal sequence". Then $F_n(r_k) \to F_D(r_k)$, $r_k \in D$. Since F_n is one of the original F_n's we have $F_n(x) \leq F_n(y)$ if $x \leq y$, so $F_D(x) \leq F_D(y)$ if $x \leq y$. Define

$$F(x) = \inf\{F_D(r) : r \in D, r > x\}.$$

By construction, F is nondecreasing, so F has at most a countable set of discontinuities. If $F(x)$ approaches a limit $L > F(x)$ as $x \to x_0^+$, let $y_1 < y_2 < \cdots$ with

$F_0(y) = 0$. For $x \geq 0$ we have $F_{n_k}(x) \geq F_{n_k}(0) \geq F_{n_k}(-1)$. Thus $\lim F_{n_k}(x) \geq 1$ is a contradiction, and therefore $F(x_0) > 0$ is impossible.

Now we show that $F_{n_k}(y) \to F(y)$ at continuity points y of F. If $x < y < z$, we have $\limsup_{k\to\infty} F_{n_k}(y) \leq \limsup_{k\to\infty} F_{n_k}(z) = F(z)$. Take the infimum over $z > y$ to obtain $\limsup_{k\to\infty} F_{n_k}(y) \leq F(y)$.

If $x < y < z$, we have $F(x) \leq \liminf_{k\to\infty} F_{n_k}(y) \leq \limsup_{k\to\infty} F_{n_k}(y) \leq F(z)$. Let $z \to y$ to obtain $F(y-) \leq \liminf_{k\to\infty} F_{n_k}(y) \leq \limsup_{k\to\infty} F_{n_k}(y) \leq F(y)$. Thus if F is continuous at y, we have $F_{n_k}(y) \to F(y)$. ∎

It must be emphasized that Helly's theorem does not say that $F_n(x) \to F(x)$. Recall that $F(x) = \lim_{y\downarrow x} F(y)$, so the discussion after (4.2) implies that every F_n is the distribution function of a random variable, yet the $F_n(x)$ for all n is possible for $F(x)$ to be strictly less than 1. (See the example that follows (7.2).)

7.2.1 *Comments.* The proof of assertion (iv) $\Rightarrow$ (i) is that $F_n(x_0) \to A$ for all x_0 we assume that $F_n(x_0) \to F_\infty(x_0) \leq M < \infty$ for all x_0, and 7.2.1 implies that there is a distribution function F and a subsequence $\{F_{n_k}\}$ with $F_{n_k}(x_0) \to F(x_0)$ for all x_0 at which F is continuous. To see this, consider $G_n(x) = 1 - F_n(x) + F_n(-x-)$.

If $F_1, F_2, \ldots$ are distribution functions on $\mathbb{R}^1$ (assumed nondecreasing in each coordinate and right-continuous) and $\lim_{n\to\infty} F_n(x) = M < \infty$ for each x, then there is a distribution function F and a subsequence $\{F_{n_k}\}$ converging to F at continuity points of F (see (7.2)). The proof is essentially the same as above.

Example (7.2) is typical in connection with the requirement of convergence of distribution functions with no weakly convergent subsequence. For example, let F_n be the distribution function of a random variable that is identically n; $F_n(x) = 0$ if $x < n$, $F_n(x) = 1$ if $x \geq n$. If F_{n_k} converges completely to F, then $F(x)$ cannot be 0 for all x or 1 for all x; hence $F(-\infty) = F(\infty) = 0$. But $F(-\infty) = \lim_k F_{n_k}(-\infty) = 1$, a contradiction.

A bounded sequence of distribution functions will always have a weakly convergent subsequence (take $M = 1$ in the above example), but such subsequences need not converge to a distribution function. Weak convergence requires $F(-\infty) = 0$ and $F(\infty) = 1$—see Examples 4.5 and 4.6. We now introduce a condition that guarantees that mass does not escape.

7.2.2 *Definition.* Let $\mathcal{M} = \{\mu_t : t \in T\}$ be a family of probability measures on the Borel sets of a metric space S. We say that $\mathcal{M}$ is *tight* if for each $\varepsilon > 0$, there is a compact set $K = K_\varepsilon \subset S$ such that $\mu_t(K) \geq 1 - \varepsilon$ for all $t \in T$. If $S = \mathbb{R}^1$, the compact sets can be replaced by finite intervals. We say that $\mathcal{M}$ is *relatively compact* if every sequence in $\mathcal{M}$ has a subsequence converging weakly to a probability measure on S.

If $\{F_\alpha : \alpha \in I\}$ is a family of distribution functions, tightness of $\{F_\alpha\}$ means tightness of the associated family of measures, and similarly for relative sequential compactness.

We may now prove the main result.

7.2.4 *Prokhorov's Theorem.* Let $\mathscr{F} = \{F_\alpha : \alpha \in I\}$ be a family of distribution functions on $\mathbb{R}$, and assume that $F_\alpha(\{x\}) = F_\alpha(\{x\}^-)$ for all x and all α. Then the family is tight if and only if it is relatively sequentially compact.

Proof. Assume $\mathscr{F}$ is tight, and let $\{F_n : n \in \mathbb{N}\}$ be a sequence from $\mathscr{F}$. By 7.2.1 and 7.2.2, there is a distribution function F and a subsequence $\{F_{n_k}\}$ such that $F_{n_k}(x) \to F(x)$, $x \in D$, for all $x \in \mathbb{R}$ and all $\omega \in \Omega$, where D is continuous. Given $x \in \mathbb{R}$, let $\{x_j\}$ and $\{y_j\}$ be points of D such that $F_\alpha(\mathbb{R} - (a, b]) < \varepsilon$ whenever $\alpha \in I$ and $b \geq x_j$ or $a \leq y_j$. If $x < x_j$ and x is a continuity point of F, such that $F_{n_k}(x) \to F(x)$ and $F_{n_k}(x_j) \to F(x_j)$, then

$$F_{n_k}(x) - F_{n_k}(y_j) = F_{n_k}(x) - F_{n_k}(x_j) + F_{n_k}(x_j) - F_{n_k}(y_j).$$

But $F_{n_k}(x) - F_{n_k}(x_j) \leq \varepsilon$ for all k, and $F_{n_k}(x_j) - F_{n_k}(y_j) \to F(x_j) - F(y_j)$ as $k \to \infty$, which follows from the definition of D. Therefore, for sufficiently large k, $F_{n_k}(x) - F_{n_k}(y_j)$ differs from $F(x) - F(y_j)$ by less than 2ε. Consequently, $F_{n_k}(x) \leq F_{n_k}(y_j) + 2\varepsilon \leq F(x_j) + 3\varepsilon$, and similarly $F_{n_k}(x) \geq F_{n_k}(x_j) - 2\varepsilon \geq F(y_j) - 3\varepsilon$. Thus $F_{n_k} \Longrightarrow F$, proving relative compactness.

Since it is metrizable, compactness is equivalent to sequential compactness. To establish the converse, suppose $\{F_\alpha : \alpha \in I\}$ is not tight. Then there is an $\varepsilon > 0$ such that for each compact K there is an F_α with $F_\alpha(K) < 1 - \varepsilon$. If $\{F_n\}$ is a subsequence converging weakly to F, then since $F = 1 - \mu_F$ is closed,

$$\liminf_n F_n(K) = \liminf_n F_n(K \cap F) \geq F(K \cap F) = F(K) < 1 - \varepsilon.$$

Thus $F_n(K) < 1 - \varepsilon$ infinitely often for all K, and it follows that $\{F_n\}$ has no vaguely convergent subsequence. ∎

Prokhorov's theorem yields the following result for sequences of random variables. If for each $n = 1, 2, \ldots$, X_n is a random variable with distribution function F_n, and the sequence $\{F_n\}$ is tight, then there is a subsequence $\{F_{n_k}\}$ and a random variable X (possibly defined on a different probability space) with distribution function F, such that F_{n_k} converges weakly to F.

Prokhorov's theorem holds equally well for distribution functions on $\mathbb{R}^k$ (see Problem 7.6), and the proof is essentially the same as above.

7.2.5 *Corollary.* Let $\{F_n\}$ be a bounded sequence of distribution functions on $\mathbb{R}^k$. If $\{F_n\}$ is tight and every weakly convergent subsequence of $\{F_n\}$ converges to the distribution function F, then $F_n \xrightarrow{\mathscr{L}} F$.

Proof. If $\int_{\mathbb{R}^k} f\,dF_n \to \int_{\mathbb{R}^k} f\,dF$ for all $f \in C_B(\mathbb{R}^k)$, then $\int_B f\,dF_n \to \int_B f\,dF$ for all $f \in C(B)$ for some bounded continuous F. Hence there is an $M > 0$ such that

$$\int_{\mathbb{R}^k} f(x)\,dF(x) - \int_{\mathbb{R}^k} f(x)\,dF_n(x) \le \varepsilon$$

for sufficiently large n, say $n \ge N$. By 7.2.1, $\{F_n : n \ge 1\}$ has a subsequence $\{F_{n_j}\}$ converging weakly to a distribution function G, and $G = F$ by hypothesis. But then $\int_{\mathbb{R}^k} f\,dF_{n_j} \to \int_{\mathbb{R}^k} f\,dF$, a contradiction. $\square$

7.2.6 Corollary. Let $\{F_n\}$ be a bounded sequence of distribution functions on $\mathbb{R}^k$. Then $\{F_n\}$ converges weakly if $\int_{\mathbb{R}^k} e^{it\cdot x}\,dF_n(x)$ approaches a finite limit as $n \to \infty$ for each $t \in \mathbb{R}^k$.

Proof. Assume $\int_{\mathbb{R}^k} e^{it\cdot x}\,dF_n(x)$ has a finite limit for all t. By 7.2.4, there is a subsequence $\{F_{n_j}\}$ converging weakly to a distribution function F. If F_n does not converge weakly to F, by 7.2.5 there is a subsequence $\{F_{n'_j}\}$ converging weakly to a distribution function $G \ne F$. We know by hypothesis that $\int_{\mathbb{R}^k} e^{it\cdot x}\,dF_{n_j}(x)$ and $\int_{\mathbb{R}^k} e^{it\cdot x}\,dF_{n'_j}(x)$ have the same limit as $j \to \infty$. Therefore $\int_{\mathbb{R}^k} e^{it\cdot x}\,dF(x) = \int_{\mathbb{R}^k} e^{it\cdot x}\,dG(x)$ for all $t \in \mathbb{R}^k$. By 7.1.4, $F = G$, a contradiction. Thus $F_n \to F$. The converse follows from the definition of weak convergence. The proof is in $\mathbb{R}^k$; distribution functions on $\mathbb{R}^k$ are characterized by their characteristic functions.

One more result is needed before we can return to weak convergence. It is a consequence of characteristic functions.

7.2.7 Inversion Inequality. Let F be a bounded distribution function on $\mathbb{R}^k$, with characteristic function ϕ. Then if B is the cube $\{x : |x_j| \le a,\ 1 \le j \le k\}$ and $\lambda > 0$,

$$\int_{|x_1| \le a, \dots, |x_k| \le a} dF(x) \le \frac{\lambda^k}{\pi^k} \int_B |\phi(t)|\,dt + 8k\lambda^{k-1} \sum_{j=1}^{k} \int_{|x_j| > \lambda} dF(x).$$

Proof.

$$\begin{aligned}
\frac{1}{\pi^k} \int_B |\phi(t)|\,dt &= \frac{1}{\pi^k} \int_{-a}^{a} \cdots \int_{-a}^{a} \left| \int_{\mathbb{R}^k} e^{it\cdot x}\,dF(x) \right| dt \\
&= \int_{\mathbb{R}^k} \left| \frac{1}{\pi^k} \int_{-a}^{a} \cdots \int_{-a}^{a} e^{it\cdot x}\,dt \right| dF(x) \\
&\quad \text{by Tonelli's theorem} \\
&= \int_{\mathbb{R}^k} \prod_{j=1}^{k} \left| \frac{\sin ax_j}{\pi x_j} \right| dF(x) \\
&\ge \inf_{|x_j| \le \lambda} \prod_{j=1}^{k} \left| \frac{\sin ax_j}{\pi x_j} \right| \int_{|x_j| \le \lambda} dF(x) \\
&= \left(\frac{2}{\pi} \int_0^{a\lambda} \frac{\sin x}{x}\,dx \right)^k \int_{|x_j| \le \lambda} dF(x).
\end{aligned}$$

In fact,

$$\lim_{t\to 0}\left(1-\frac{\sin t}{t}\right)=\qquad \text{and } t\neq\frac{\pi}{2}$$

so we may take $k=2$.

7.2.8 Lévy's Theorem. Let $\{F_n\}$ be a bounded sequence of distribution functions on $\mathbb{R}$, and let $\{\phi_n\}$ be the corresponding sequence of characteristic functions: $\phi_n(t)=\int e^{itx}\,dF_n(x)$. If $F_n\xrightarrow{w}F$, where F is a distribution function with characteristic function ϕ, then $\phi_n(t)\to\phi(t)$ for all t. Conversely, if ϕ_n converges pointwise to a complex-valued function ϕ which is continuous at $t=0$, then ϕ is the characteristic function of some bounded nondecreasing function F, and $F_n\xrightarrow{w}F$.

PROOF. The first assertion follows from the definition of weak convergence, since $\cos tx$ and $\sin tx$ are bounded and continuous in x for all t, with a discontinuity at the origin. We claim that $\{F_n\}$ is tight. Let $\varepsilon>0$.

$$\int_{-A}^{A}\left(1-\frac{\sin ux}{ux}\right)dF_n(x)\geq\frac{1}{u}\int_{-u}^{u}\left[\phi_n(t)-\operatorname{Re}\phi_n(t)\right]dt \qquad u>0$$

$$\geq\frac{2}{u}\int_{0}^{u}\left[\phi_n(t)-\operatorname{Re}\phi(t)\right]dt \qquad \text{as} \qquad n\to\infty$$

by the dominated convergence theorem.

Since ϕ is continuous at 0,

$$\frac{2}{u}\int_{0}^{u}\left[\phi(t)-\operatorname{Re}\phi(t)\right]dt=0 \qquad \text{as} \qquad u\to 0.$$

Hence, given $\varepsilon>0$, we may choose u so small that

$$\int_{-A}^{A}\left(1-\frac{\sin ux}{ux}\right)dF_n(x)<\varepsilon \qquad \text{for all} \qquad n$$

proving tightness. By 7.2.6, F_n converges weakly to a distribution function F, hence ϕ_n converges pointwise to the characteristic function of F. But we know that $\phi_n\to\phi$, so that ϕ is the characteristic function of F. □

This converse version of Lévy's theorem is often useful.

7.2.5 Theorem. Let $\{F_n\}$ be a bounded sequence of distribution functions, and $\{f_n\}$ the corresponding sequence of characteristic functions. If F is a bounded distribution function with characteristic function f, then $\lim_{n\to\infty} f_n(t) = f(t)$ for all t if and only if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x at which F is continuous. In this case, F_n converges to F uniformly on bounded intervals.

Proof. If $F_n \to F$ then $f_n(t) \to f(t)$ for all t by definition of weak convergence. If $f_n(t) \to f(t)$ for all t, then by 7.2.4, F_n converges weakly to the distribution function whose characteristic function is f, namely F.

Now let I be a bounded interval in $\mathbb{R}$. Then

$$\sup_{x\in I} |F_n(x) - F(x)| \le \int_{-T}^{T} |e^{itx} f_n(t) - e^{itx} f(t)| \, d\mu_g(t),$$

hence

$$\sup_{x\in I} |F_n(x) - F(x)| \le \int_{-T}^{T} |f_n(t) - f(t)| \, d\mu_g(t) + \varepsilon \mu_g(\mathbb{R}) = \varepsilon.$$

Since $F_n \to F$, $\{F_n\}$ is relatively compact, and therefore tight by 7.2.4. Thus if $\varepsilon > 0$ is given, we may choose T so that $2F_n(-T) < \varepsilon/2$ for all n. If $x \in \mathbb{R}$, set $I = [x, x+1]$ and $\{x_1, \dots, x_k\} = \{x, x+1\}$. Then

$$\sup_{x\in I} |F_n(x) - F(x)| \le \sup_{1\le i\le k} |F_n(x_i) - F(x_i)| + \sum_{i=1}^{k} |F(x_i) - F(x_i - 0)|$$

$$< \varepsilon \quad \text{for all sufficiently large } n, \text{ since } \varepsilon > 0 \text{ was arbitrary.}$$

Thus the F_n are equicontinuous on $\mathbb{R}$. But equicontinuity and pointwise convergence imply uniform convergence on compact sets (see Ash, 1993, 7.4, Problem 9); hence $F_n \to F$ uniformly on I. $\square$

Problems

1. If X has density $\frac{1}{2}\exp(-|x|)$, the characteristic function of X is $\phi(t) = 1/(1+t^2)$. If $\alpha > 0$, $\alpha \neq 1$, then $|\phi(t)|^\alpha = \phi(\beta t)$ for some $\beta > 0$. (See Ash, 1990, 7.1.65.) However, by 8.1 it is possible to construct a different characteristic function f that agrees with ϕ on $[-1, 1]$, as follows.

   Show that the Fourier series corresponding to the interval $[-1, 1]$ of $f(x) = \phi(x)$ [x extended periodically] is

$$f(x) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{2}{\pi^2(2n-1)^2} \exp(i(2n-1)\pi x).$$

Since F is continuous and of bounded variation, and $\int e^{itx}\,dF(x) = f(t)$, the Stieltjes integral converges uniformly in t; see Titchmarsh [1939], p. 411]. Thus F is a continuous uniform mixture with $P\{X = k\} = \frac{1}{2}$, $\mu = \frac{1}{2}(a+b)$, $\sigma^2 = \frac{1}{12}(b-a)^2 = \frac{1}{3}(\frac{1}{2}(b-a))^2 + \frac{1}{4}(a+b)^2$. If a is an integer, then X has characteristic function f. Since f is periodic, we have $f \neq \hat{n}$, as desired.

4 Using Problem 3, give examples to show that the following results are impossible (where $\hat{n}$ and $\hat{m}$ are characteristic functions of random variables):

(a) $\hat{n}(t) = \hat{n}(ct)$ does not imply $c = \pm 1$.

(b) $\hat{n}(t) = \hat{m}(t)$ does not imply $n = m$ or $n = -m$.

5. Give an example of a sequence of characteristic functions converging to a function that is not a characteristic function.

6 (a) If $\{p_j\}$ is a sequence of nonnegative numbers not exceeding 1 (where $\sum p_j$ may diverge), show that $\prod_{j=1}^{n}(1 - p_j + p_j e^{it})$ converges for all t if and only if $\sum_{j=1}^{\infty} p_j$ converges.

(b) If $X_1, X_2, \ldots$ are independent random variables, show that the sequence of characteristic functions of $S_n = X_1 + \cdots + X_n$ converges if and only if the corresponding sequence of distribution functions has a weakly convergent subsequence. (This was verified directly in Problem 2.)

7 Let $F_1, F_2, \ldots$ be distribution functions on $\mathbb{R}$, and assume $F_n(-\infty) = 0$, $n = 1, 2, \ldots$, $F_n(\infty) = 1$, $n = 1, 2, \ldots$. Give examples to show that the following situations are possible.

(a) $F_n(x) \to F(x)$ for all x, F is continuous, but F_n does not converge weakly to F.

(b) $F_n(x) - F_n(x-) \to F(x) - F(x-)$ for all x, F is continuous, but $\lim_{n\to\infty} F_n(x)$ exists and is not $F(x)$ for some x; in other words, F_n does not converge weakly to F.

8 Let $F_1, F_2, \ldots$ be distribution functions on $\mathbb{R}$, and assume that $F_n(x) \to F(x)$ for all x and $F_n(a,b] \to F(a,b]$ for all $a < b$, $a, b \in C$, where C is dense and F is continuous. Show that F_n converges weakly to F iff $\{F_n\}$ is tight.

9. Let $F_1, F_2, F_3, \ldots$ be distribution functions on $\mathbb{R}$, and assume that $F_n(x) = F_n(x-)$ for all $n \geq 1$ and $x \in \mathbb{R}$. Show that F_n converges weakly to F iff $F_n(a,b] \to F(a,b]$ for all $a < b$, $a, b \in \mathbb{R}$ at which F is continuous, and $\int g(x)\,dF_n(x) \to \int g(x)\,dF(x)$.

10. Let $Y_1, Y_2, \ldots$ be independent random variables, and let $S_n = Y_1 + \cdots + Y_n$. If $\hat{n}_j$ is the characteristic function of Y_j and

n is a fixed real number such that $\lambda_k(\alpha) \neq 0$, $k = 1, 2, \ldots$. Prove that

$$\lambda_\alpha = \prod_{k=1}^{\infty} \left[\frac{\lambda_k}{\lambda_k(\alpha)} \right]^{\alpha_k} \exp[\beta(\alpha)]$$

(a) Show that $\{X_n - \beta(\alpha)\}$ converges in probability to X_α if and only if $\{\log X_n\}$ converges in probability to $\log X_\alpha$. (In other words, $\log X_n \xrightarrow{p} \log X_\alpha$ if and only if $X_n \xrightarrow{p} X_\alpha$.)

(b) Assume that $\sum_{k=1}^{\infty} \alpha_k < \infty$. Show that there is an open interval $I \subset \mathbb{R}$ with $1 \in I$, such that for each $\alpha \in I$, $\exp[\alpha \sum_{k=1}^{n} X_k(\omega)]$ converges for almost every ω.

(c) Under the hypothesis of part (b), show that for almost every ω, $\exp[\alpha \sum_{k=1}^{n} X_k(\omega)]$ can converge for at most one $\alpha \in I$. (Use analytic continuation.) Thus by Problem 4, $\sum_{k=1}^{\infty} X_k(\omega)$ converges to a finite limit.

(d) Conclude that for a series of independent random variables, convergence in distribution, convergence in probability, and convergence almost everywhere are equivalent.

7.3 Convergence to a Normal Distribution

Let $X_1, X_2, \ldots$ be an independent sequence of random variables, with each X_k having finite mean μ_k and finite variance σ_k^2. Let $S_n = \sum_{k=1}^{n} X_k$, $n = 1, 2, \ldots$; then $E(S_n) = \sum_{k=1}^{n} \mu_k$, $\operatorname{var} S_n = \sigma_n^2 = \sum_{k=1}^{n} \sigma_k^2$. We consider the normalized sum $T_n = \sigma_n^{-1}[S_n - E(S_n)]$, which has mean 0 and variance 1. To avoid degenerate cases we assume that $\sigma_n > 0$ for sufficiently large n. Under the Lindeberg hypothesis (7.3.1), it will turn out that T_n converges in distribution to $N(0, 1)$ as $n \to \infty$.

Let X be a random variable having the normal distribution with mean 0 and variance 1, so that the distribution function of X is

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{y^2}{2}\right] dy$$

We ask for conditions under which T_n converges in distribution to X.

A large number of preliminary results were derived by Lindeberg and Feller. The final result was obtained by Feller, giving conditions that are sufficient and "almost" necessary for convergence to X. We begin with Lindeberg's basic result.

7.3.1 Lindeberg's Theorem. Let $S_n = X_1 + \cdots + X_n$, $n = 1, 2, \ldots$, where the X_k are independent random variables with finite mean μ_k and finite

variances σ_n^2. Let $T_n = \sigma_n^{-1}(S_n - E(S_n))$, where $\sigma_n^2 = \mathrm{Var}\, S_n = \sum_{k=1}^n \sigma_k^2$, and let F_n be the distribution function of T_n. If for every $\tau > 0$

$$\frac{1}{\sigma_n^2} \sum_{k=1}^{n} \int_{\{|x-m_k| \geq \tau \sigma_n\}} (x-m_k)^2 \, dF_k(x) \to 0 \quad \text{as} \quad n \to \infty,$$

then T_n converges in distribution to a random variable T^* that is normal with mean 0 and variance 1.

Before proving the theorem, we examine some of its implications. Lindeberg's theorem implies that $T_n \xrightarrow{\mathcal{L}} N^*$ under any one of the following conditions:

1. *The uniformly bounded case.* Assume $|X_k - m_k| \leq M$ for all k, and $\sigma_n \to \infty$. Then

$$\int_{\{|x-m_k| \geq \tau \sigma_n\}} (x-m_k)^2 \, dF_k(x) = E[(X_k - m_k)^2 I_{\{|X_k - m_k| \geq \tau \sigma_n\}}]$$

$$\leq (2M)^2 P\{|X_k - m_k| \geq \tau \sigma_n\}$$

$$\leq \frac{(2M)^2 \sigma_k^2}{\tau^2 \sigma_n^2}$$

by Chebyshev's inequality. Thus

$$\frac{1}{\sigma_n^2} \sum_{k=1}^{n} \int_{\{|x-m_k| \geq \tau \sigma_n\}} (x-m_k)^2 \, dF_k(x) \leq \frac{(2M)^2}{\tau^2 \sigma_n^2} \to 0.$$

2. *The identically distributed case.* Assume that the X_k are all with finite mean m and finite variance $\sigma^2 > 0$. If F is the distribution function of the X_k, then

$$\frac{1}{\sigma_n^2} \sum_{k=1}^{n} \int_{\{|x-m_k| \geq \tau \sigma_n\}} (x-m_k)^2 \, dF_k(x)$$

$$= \frac{1}{n\sigma^2} \sum_{k=1}^{n} \int_{\{|x-m| \geq \tau \sqrt{n}\sigma\}} (x-m)^2 \, dF(x)$$

$$= \sigma^2 \int_{\{|x-m| \geq \tau \sqrt{n}\sigma\}} (x-m)^2 \, dF(x)$$

$$\to 0 \quad \text{since } \sigma^2 \text{ is finite and } \{|x-m| \geq \tau\sqrt{n}\sigma\} \downarrow \emptyset \text{ as } n \to \infty.$$

3. *The Bernoulli case.* Let S_n be the number of successes in a Bernoulli trials, with probability of success p on a given trial. We may write

$$S_n = X_1 + \cdots + X_n,$$

where the X_i are independent and $P[X_i = 1] = p, P[X_i = 0] = q = 1 - p$. (We may take X_i as the indicator of a success on trial i.) The weak law applies. Here $m = E[X_i] = p, \sigma^2 = \mathrm{Var}[X_i] = E[X_i^2] - p^2 = p - p^2 = pq$, and $\mathrm{Var}[S_n] = npq, \sigma_n^2 = \mathrm{Var}^2 = npq$. Thus

$$T_n = \frac{S_n - np}{(npq)^{1/2}}$$

and $T_n = \frac{1}{\sigma_n} X^*$, that is, $P[T_n \le x] = F^*(x)$ for all x.

4. *Lyapunov condition.* Assume that

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - m_k|^{2+\delta}] \to 0 \quad \text{for some } \delta > 0.$$

Then

$$\begin{aligned} \int_{|x-m_k| \ge \varepsilon s_n} (x - m_k)^2 \, dF_k(x) &\le \int_{|x-m_k| \ge \varepsilon s_n} |x - m_k|^{2+\delta} (\varepsilon s_n)^{-\delta} \, dF_k(x) \\ &\le (\varepsilon s_n)^{-\delta} \int_{-\infty}^{\infty} |x - m_k|^{2+\delta} \, dF_k(x) \end{aligned}$$

Thus

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \int_{|x-m_k| \ge \varepsilon s_n} (x - m_k)^2 \, dF_k(x) \le \frac{1}{\varepsilon^{\delta} s_n^{2+\delta}} \sum_{k=1}^{n} E[|X_k - m_k|^{2+\delta}] \to 0$$

Proof of Theorem 7.1.1. We may assume without loss of generality that $m_1 = 0$. For $n \geq 1$ we have proved the theorem under the restriction $\sigma_n^2 > 0$. Now we have proved the theorem under the restriction $\sigma_n^2 > 0$. Let

$$S_n' = X_n' \qquad \text{for } n = 1, 2, \ldots$$

$$S_n' = \sum_{i=1}^{n} X_i'$$

$$T_n' = \frac{S_n'}{\sigma_n'} = \frac{S_n}{\sigma_n} = T_n$$

Since

$$\int_{|x| \geq \varepsilon \sigma_n'} x^2 \, dF_{X_n'}(x) = \int_{|x| \geq \varepsilon \sigma_n} x^2 \, dF_{X_n}(x)$$

the Lindeberg hypothesis applies to the random variables X_n' because $\sigma_n'^2 = \sigma_n^2$. Hence $T_n' \to N(0,1)$ as desired.

The following formulas will be needed. If y is any real number and z a complex number with $|z| \leq \frac{1}{2}$,

$$e^{iy} = 1 + iy + \frac{\theta y^2}{2}, \tag{1}$$

$$e^{iy} = 1 + iy - \frac{y^2}{2} + \frac{\theta |y|^3}{6}, \tag{2}$$

where θ and θ depend on y and $|\theta| \leq 1$, $|\theta| \leq 1$,

$$\operatorname{Log}(1+z) = z - \frac{z^2}{2} + \theta |z|^3, \tag{3}$$

where "Log" denotes the principal branch of the logarithm and $|\theta| \leq 1$, θ depending on z. (These formulas are exercises in calculus. For more details see Feller [1971, p. 172].)

Throughout the proof, φ_n will denote the characteristic function of X_n, and t will be a fixed real number. The characteristic function of $T_n = S_n/\sigma_n$ is

$$\psi_n(t) = \varphi_1\left(\frac{t}{\sigma_n}\right) \varphi_2\left(\frac{t}{\sigma_n}\right) \cdots \varphi_n\left(\frac{t}{\sigma_n}\right) = \prod_{k=1}^{n} \varphi_k\left(\frac{t}{\sigma_n}\right) \tag{4}$$

By (1) and (2),

$$\begin{aligned}\Delta(x) = &\int_{-\infty}^{\infty} e^{ixy}\,dH(y)\\ &- \int_{|y|\le\varepsilon\sqrt{n}} \left(1 + ixy - \frac{x^2y^2}{2}\right) dF(y)\\ &\int_{|y|>\varepsilon\sqrt{n}} \left(1 + ixy - \frac{x^2y^2}{2} + \frac{|x|^3|y|^3}{6}\right) dF(y).\end{aligned}$$

$|\theta_1|, |\theta_2| \le 1$, θ depending on x and θ now:

$$\int_{-\infty}^{\infty} ixy\,dF(y) = ixE(X) = 0,$$

we may drop the same terms ixy, too. Thus

$$\begin{aligned}\Delta(x) = &\frac{x^2}{2}\int_{|y|\le\varepsilon\sqrt{n}} \theta_1 y^2\,dF(y)\\ &\frac{x^2}{2}\int_{|y|>\varepsilon\sqrt{n}} y^2\,dF(y) \qquad \frac{|x|^3}{6}\int_{-\infty}^{\infty} \theta_2 |y|^3\,dF(y). \qquad (2)\end{aligned}$$

Now

$$\left|\frac{1}{2}\int_{|y|\le\varepsilon\sqrt{n}} \theta x^2 y^2\,dF(y)\right| \le \frac{1}{2}\int_{-\infty}^{\infty} x^2 y^2\,dF(y),$$

hence

$$\frac{1}{2}\int_{-\infty}^{\infty} \theta x^2 y^2\,dF(y) = \theta_3 \int_{-\infty}^{\infty} x^2 y^2\,dF(y) \qquad \text{where} \qquad |\theta_3| \le \frac{1}{2}.$$

Similarly

$$\begin{aligned}\frac{1}{6}\int_{|y|>\varepsilon\sqrt{n}} \theta_2 |y|^3\,dF(y)\Big| &\le \frac{1}{6}\int_{|y|>\varepsilon\sqrt{n}} |y|^3\,dF(y)\\ &\le \frac{1}{6}\int_{|y|>\varepsilon\sqrt{n}} \frac{y^2}{\varepsilon\sqrt{n}}\,|y|^3\,dF(y)\end{aligned}$$

since $|y| > \varepsilon\sqrt{n}$ implies $\frac{|y|}{\varepsilon\sqrt{n}} > 1$. Hence

$$\frac{1}{6}\int_{|y|>\varepsilon\sqrt{n}} \theta_2 |y|^3\,dF(y) = \theta_4 \int_{|y|>\varepsilon\sqrt{n}} \varepsilon\sqrt{n}\,y^2\,dF(y) \qquad |\theta_4| \le \frac{1}{6}.$$

Thus if we set

$$a_{nk} = \frac{1}{s_n^2} \int_{|x|\le \varepsilon s_n} x^2\, dF_k(x),$$

$$\beta_{nk} = \frac{1}{s_n^2} \int_{|x|>\varepsilon s_n} x^2\, dF_k(x) \qquad (k=1,\dots,n),$$

(5) becomes

$$\Delta_n\left(\frac{u}{s_n}\right) = \prod_{k=1}^{n} \gamma_{nk}, \tag{6}$$

where

$$\gamma_{nk} = \varphi_k\left(\frac{u}{s_n}\right) e^{\frac{u^2}{2s_n^2}\sigma_{nk}^2} = e^{\frac{u^2}{2s_n^2}\sigma_{nk}^2}\left(1 + \theta_{nk}\beta_{nk}\right). \tag{7}$$

Now $\sum_{k=1}^{n} \sigma_{nk}^2 \to 1$ by hypothesis, and

$$\sum_{k=1}^{n} (a_{nk} + \beta_{nk}) = \frac{1}{s_n^2} \sum_{k=1}^{n} \sigma_k^2 = 1.$$

Thus by (7) and the fact that $|\theta_{nk}| \le 1$, we have

$$\max_{1\le k\le n} \sigma_{nk}^2 \le \varepsilon^2 + \max_{1\le k\le n} a_{nk} \le \varepsilon^2 + \sum_{k=1}^{n} a_{nk} \le \varepsilon^2 + \frac{\sigma^2}{s_n^2} \to \varepsilon^2,$$

and consequently, for sufficiently large n, by (7),

$$\frac{u^2}{2} - \sum_{k=1}^{n} \log|\gamma_{nk}| = \frac{u^2}{2} - \sum_{k=1}^{n} \left(\sigma_{nk}^2 - \theta_{nk}^2\sigma_{nk}^4\right) \qquad |\theta| \le 1.$$

Now

$$\sum_{k=1}^{n} \sigma_{nk}^4 \le \left(\max_{1\le k\le n} \sigma_{nk}^2\right) \sum_{k=1}^{n} \sigma_{nk}^2, \tag{8}$$

and it follows from (8) that if $\varepsilon > 0$ is given, and n is chosen sufficiently large, depending on ε and σ, then

$$\left|\frac{u^2}{2} + \sum_{k=1}^{n} \log|\gamma_{nk}|\right|$$

will be less than ε for sufficiently large n. Since the exponential function is continuous, we obtain from (i) that

$$\exp\left[\frac{it}{\sigma_n}\sum_{k=1}^{n}\mu_k\right]\prod_{k=1}^{n}\varphi_k\left(\frac{t}{\sigma_n}\right)\to\varphi\left(\frac{t}{\sigma}\right),$$

in the sense of [see (4)], $\varphi_{S_n^*}(t)\to\exp(-t^2/2)$, the characteristic function of $\mathcal{N}(0,1)$. By 7.2.8, $S_n^*\xrightarrow{\mathcal{D}}\mathcal{N}(0,1)$. □

There is another proof of the Lindeberg theorem that does not work with characteristic functions but rather with Taylor's theorem; see Billingsley (1968, p. 49).

We now show that the Lindeberg hypothesis is not necessary for convergence to $\mathcal{N}$. If the Lindeberg condition holds, we obtain that for all $\varepsilon>0$

$$P\left\{\max_{1\le k\le n}\frac{|X_k-\mu_k|}{\sigma_n}\ge\varepsilon\right\}\to 0\qquad\text{as}\qquad n\to\infty.$$

(why?). This is another way of saying that the contribution of each individual random variable $(X_k-\mu_k)/\sigma_n$ to the sum $S_n^*=\sum_{k=1}^{n}(X_k-\mu_k)/\sigma_n$ is small relative to the sum itself.

If we can construct an example of a sequence that is not tight but for which $S_n^*\xrightarrow{\mathcal{D}}\mathcal{N}$, we can have convergence to $\mathcal{N}$ without the Lindeberg condition. Let $Y_1,Y_2,\ldots$ be independent, normally distributed random variables with $EY_k=0$ and $\operatorname{Var}Y_k=2^{k-2}$, $k=2,3,\ldots$, $\operatorname{Var}Y_1=1$. Then $\sigma_n^2=\sum_{k=1}^{n}\sigma_k^2=2^{n-1}$, hence $S_n=\sum_{k=1}^{n}Y_k$ is normal with mean 0 and variance $\frac{1}{2}$. Therefore

$$\max_{1\le k\le n}P\left\{\left|\frac{Y_k}{\sigma_n}\right|>\varepsilon\right\}=P\left\{\frac{Y_n}{\sigma_n}>\varepsilon\right\},$$

where ε is a positive constant not depending on n. Thus the Y_k/σ_n are not tight, although $S_n^*\xrightarrow{\mathcal{D}}\mathcal{N}$. To see this, we must show $P\{Y_n/\sigma_n>\varepsilon\}\not\to 0$.

If we impose the additional requirement that the Lindeberg condition becomes necessary and sufficient for convergence to $\mathcal{N}$.

7.3.2 *Theorem.* Let $X_1, X_2, \ldots$ be independent random variables, with each X_i having finite mean μ_i and finite variance σ_i^2. Then the Lindeberg condition

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|x-\mu_i|\geq \varepsilon s_n} (x-\mu_i)^2 \, dF_{X_i}(x) = 0 \quad \text{for all } \varepsilon > 0$$

holds if and only if $\frac{S_n - m_n}{s_n} \xrightarrow{\mathcal{L}} N$ and the $\max_{1\leq i\leq n} \frac{\sigma_i}{s_n} \to 0$.

Proof. [illegible] The "only if" part follows from 7.3.1 and the above remarks. The "if" part, due to Feller (1968), is rather lengthy and is omitted in accordance with [illegible].

In the Lindeberg theorem we have normalized the sum S_n in a special way, that is, we have considered a sequence of random variables $\frac{S_n - m_n}{s_n}$, where m_n is the mean and s_n the standard deviation of S_n. We might ask whether a different choice of constants a_n and b_n would produce different results, for example, convergence to a nonnormal limiting variable. One form of this answer may be found by the "theorem on convergence of types," which we now develop.

7.3.3 *Definition.* Two random variables X and Y (or their distribution functions F_X and F_Y) are said to be of the same positive type iff X and $aX + b$ have the same distribution for some $a > 0$ and b, that is, $F_X(x) = F_{aX+b}(x)$ for all $x \in \mathbb{R}$ and X are of the same type iff X and $aX + b$ have the same distribution for some $a \neq 0$ and b. (Note that $X \stackrel{\mathcal{L}}{=} -X$ indicates that X and Y have the same distribution.)

The notion of type is preserved under convergence in distribution as the following basic theorem shows.

7.3.4 *Convergence of Types Theorem.* (a) Let $X_n \xrightarrow{\mathcal{L}} X$, $Y_n \xrightarrow{\mathcal{L}} Y$, where for each n, X_n and Y_n are of the same positive type, with $X_n \stackrel{\mathcal{L}}{=} a_n^{-1}(Y_n - b_n)$, $a_n > 0$. Assume X and Y are nondegenerate, that is, not constant. Then there are real numbers a and b, with $a > 0$, such that $a_n \to a$, $a_n b_n \to b$, and $Y \stackrel{\mathcal{L}}{=} a^{-1}(Y - b)$, that is, X and Y are of the same positive type.

(b) If $X_n \xrightarrow{\mathcal{L}} X$, $Y_n \xrightarrow{\mathcal{L}} Y$, where for each n, X_n and Y_n are of the same type, with $X_n \stackrel{\mathcal{L}}{=} a_n^{-1}(Y_n - b_n)$, $a_n > 0$, and X and Y are nondegenerate

that there are real numbers a and b, with $a \neq 0$, such that $aX + b \stackrel{D}{=} X$ and $X \stackrel{D}{=} a^{-1}(X - b)$; thus X and $aX + b$ are of the same type.

The proof is an interesting exercise in real analysis, and is given in Appendix 4.

We can now consider the question raised earlier about the normalizing constants. If $a_n^{-1}(S_n - b_n) \xrightarrow{D} Y$, and $a_n^{-1}(S_n - A_n) \xrightarrow{D} X$, where Y is normal and X is nondegenerate, then X must be normal too.

$$a_n^{-1}(S_n - b_n) = \frac{a_n}{a_{kn}}\left[a_{kn}^{-1}(S_{kn} - a_{kn}) + a_{kn}^{-1}(a_{kn} - b_{kn})\right] + a_n^{-1}(b_{kn} - b_n).$$

Hence by the convergence of types theorem, X and $a^{-1}(Y - b)$ have the same distribution for some a, b with $a \neq 0$. But a brief computation then shows that X is normal. (Thus the only stable distributions are the normal ones; see the Exercises of 4.2.4.) Thus if one set of constants produces a normal limit, all nondegenerate limits are normal.

The restriction that X be nondegenerate is necessary. For given any real number α, it is always possible to choose the constants so that $a_n^{-1}(S_n - b_n) \xrightarrow{D} \alpha$ (see Problem 2).

The most striking aspect of Theorem 7.3.2 that is worth mentioning. Under the sole hypothesis, the Lindeberg condition is necessary and sufficient for $\frac{S_n}{s_n} \xrightarrow{D} N(0,1)$. Thus if the $\sigma_j^2 = \operatorname{var} X_j$ are very unequal, the Lindeberg condition holds, $s_n^{-1}(S_n - ES_n)$ cannot converge in distribution to a normal $(0,1)$ random variable. However, it is possible for $a_n^{-1}(S_n - ES_n)$ to converge in distribution to a nondegenerate random variable X that is not normal $(0,1)$. In Problem 4, an example is given in which X is uniform with a continuous density on $(-1,1)$.

In fact, there are examples of independent (but not identically distributed) random variables $X_1, X_2, \ldots$, each with finite mean and variance, such that for some constants a_n, b_n we have $a_n^{-1}(S_n - b_n) \xrightarrow{D} X$, where X is not normal and not nondegenerate. The construction of such examples is quite elaborate, and we refer the reader to Gnedenko and Kolmogorov (1954, p. 199).

Problems

1. Let $X_1, X_2, \ldots$ be an arbitrary sequence of random variables. Given any real number α, show that constants a_n and b_n can always be chosen so that $a_n^{-1}(X_1 + \cdots + X_n - b_n)$ converges in distribution to a random variable X with $P\{X \leq \alpha\} = \frac{1}{2}$.

2. Give an example of sequences of random variables $\{X_n\}$ and $\{Y_n\}$ such that $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ for real numbers X and Y, but $X_n + Y_n \not\xrightarrow{d} X + Y$. (Hint: Let X and Y be two of the same type, but X_n has no limit.)

3. Let $\{X_{ij}, i = 1, 2, \ldots, j = 1, \ldots, n\}$ be a double sequence of random variables, and let ψ_{ij} be the characteristic function of X_{ij}. Show that the X_{ij} are uniformly infinitesimal if

$$\max_{1 \le j \le n} |\psi_{ij}(t) - 1| \to 0$$

as $i \to \infty$ for every $t \in (0, \infty)$ iff

$$\prod_{j=1}^{n} \psi_{ij}(t) \to 1$$

as $i \to \infty$, and in this case the convergence is uniform on any bounded interval. (Use 7.2.3 in the "if" part.)

4. Let $X_1, X_2, \ldots$ be independent random variables defined as follows:

$X_1 = \pm 1$ with equal probability.
If $k > 1$, then $\pm k$ is each equally likely with probability $\frac{1}{2k^2}$ and

$$P[X_k = 1] = P[X_k = -1] = \frac{1}{2k^2},$$

$$P[X_k = k] = P[X_k = -k] = \frac{1}{2k^2}\left(1 - \frac{1}{k^2}\right),$$

$$P[X_k = 0] = 1 - \frac{1}{k^2} = \frac{1}{k^2}\left(1 - \frac{1}{k^2}\right).$$

Define

$$Y_{n,r} = \begin{cases} X_r & \text{if } |X_r| \le \sqrt{n}, \\ 0 & \text{if } |X_r| > \sqrt{n}. \end{cases}$$

Establish the following:

(a) The X_k's satisfy the Lyapunov condition.

(b) The Lindeberg condition fails for the X_k, but holds for the $X_{n,r}$.

Furthermore, if $S_n = \sum_{k=1}^{n} X_k$, then $S_n/s_n \xrightarrow{d}$ normal $(0, 1)$, where $s_n^2 = \operatorname{Var} S_n = n/2$.

(c) If $X_n \to \sum_{k=1}^{n} Y_k$ then $P[|X_n| \le x] \to 0$ as $n \to \infty$.

(d) $\sqrt{n}\,\bar{X}_n/S_n \xrightarrow{d}$ normal $(0, 1)$, but $X_n/\sqrt{n} \xrightarrow{d}$ normal $(0, 1)$.

7.4 Weak Distributions

If $Y_1, Y_2, \ldots$ are independent, identically distributed random variables with finite mean μ and finite variance σ^2, we know from the previous section that $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N$ where N has the normal $(0, 1)$ distribution. The distribution of a sequence $X_n^* = (X_n - A_n)/B_n$ is not unique. If we require that X_n^* is positive we obtain a non-normal limit. For example, let the X_i have the Cauchy density $f(x) = \sigma/\pi(x^2 + \sigma^2)$, $-\infty < x < \infty$, $\sigma > 0$ a fixed positive constant. The corresponding characteristic function is $\phi(u) = e^{-\sigma|u|}$. (see Ash, 1972, p. 161, for the computation). Therefore $\bar{X}_n$ has characteristic function $[\phi(u/n)]^n = e^{-\sigma|u|}$; hence $\bar{X}_n$ has the same Cauchy distribution $[\phi(u/n)]^n = e^{-\sigma|u|}$. Thus $n^{-1}\bar{X}_n \xrightarrow{d} X$, where X has the Cauchy density with parameter σ. Since $P[|X| < \infty] = 1$, this does not contradict the previous result.

The following investigation is suggested. Let $Y_1, Y_2, \ldots$ be iid random variables. Find $a_n^{-1}(X_n - b_n) \xrightarrow{d} X$, where X has a possible distribution of the type given in Problem 7.3.6. For example, try $a_n = n^{\alpha}$ for various negative α, and a number of choices for the two subsequences corresponding to $a_n > 0$ and $a_n < 0$. In fact the possible limiting distributions may be completely characterized as follows.

7.4.1 Definition. A random variable Y (or its distribution function F) is said to be *stable* if for any $A > 0$, $B > 0$ there exist $C > 0$ and D such that if Y_1, Y_2 are iid random variables with distribution function F, then $Y = AY_1 + BY_2 + D$ has the same probability type as Y in the sense that $Y \stackrel{d}{=} C^{-1}(Y_1 + Y_2)$, or equivalently $[\phi(u)]^2 = \exp[iDu]\phi(Cu)$ for all $u \in \mathbb{R}$. For specified $a_n > 0$ and b_n, a sequence $a_n^{-1}(X_n - b_n) \xrightarrow{d} X$, where $a_n > 0$ and $X_n = \sum_{k=1}^{n} Y_{nk}$ for X_n iid, is called a *sequence of normed sums*.

7.4.2 Theorem. The random variable X is stable iff there is a sequence of normed sums converging in distribution to X.

Proof. If X is stable, let $X_1, X_2, \ldots$ be iid with $X_j \stackrel{d}{=} X$. Then $a_n^{-1}(X_1 + \cdots + X_n - b_n)$ for appropriate $a_n > 0$ and b_n; in particular $a_n^{-1}(\sum_{k=1}^{2^n} X_k - b_{2^n}) \xrightarrow{d} X$.

Assume $X_1, X_2, \ldots$ are i.i.d. with $\mu = E[X_i] = 0$, $\sigma^2 = \mathrm{Var}(X_i) < \infty$, and $M_X(t)$ is finite for $|t| < h$ for some $h > 0$. Let the positive integer r be chosen so that $r > \sigma^2$. Define

$$S_n^{(1)} = X_1 + \cdots + X_n,$$
$$S_n^{(2)} = X_{n+1} + \cdots + X_{2n},$$
$$\vdots$$
$$S_n^{(r)} = X_{(r-1)n+1} + \cdots + X_{rn}.$$

Let $c > 0$ be such that

$$\frac{S_n^{(1)}}{\sigma\sqrt{n}} + \frac{S_n^{(2)}}{\sigma\sqrt{n}} + \cdots + \frac{S_n^{(r)}}{\sigma\sqrt{n}} = Z_n^{(1)} + Z_n^{(2)} + \cdots + Z_n^{(r)},$$

where $Z_n^{(1)}, \ldots, Z_n^{(r)}$ are independent. Now $Z_n^{(i)} \xrightarrow{d} Z^{(i)}$ for all i, hence $Z_n^{(1)} + \cdots + Z_n^{(r)} \xrightarrow{d} Z^{(1)} + \cdots + Z^{(r)}$ for each r. It follows from 7.2.9 that $W_n = \frac{S_{rn}}{\sigma\sqrt{rn}} \xrightarrow{d} Z_1 + \cdots + Z_r$, where $Z_1, \ldots, Z_r$ are iid with $Z_i \stackrel{d}{=} N$. But we may also write

$$W_n = \frac{X_1 + \cdots + X_n}{\sigma\sqrt{rn}} + \frac{X_{n+1} + \cdots + X_{2n}}{\sigma\sqrt{rn}} + \cdots + \frac{X_{(r-1)n+1} + \cdots + X_{rn}}{\sigma\sqrt{rn}}$$
$$= a_n^{(1)} U_{1n} + a_n^{(2)} U_{2n} + \cdots + a_n^{(r)} U_{rn},$$

where $a_n^{(i)} = \sigma_{in}/\sigma\sqrt{r} > 0$. This suggests

$$U_{in} = \frac{S_n^{(i)}}{\sigma_{in}\sqrt{n}} \xrightarrow{d} Z_i, \quad \text{where } \sigma_{in}^2 = \mathrm{Var}(X_{(i-1)n+1} + \cdots + X_{in}).$$

By 7.2.7 (a), $a_n^{(i)}$ converges to a limit $a_i > 0$. If $S_n^{(i)}$ approximates a line in β_i, and $X \stackrel{d}{=} a_1 Z_1 + \cdots + a_r Z_r$, then X is stable. $\square$

In the above proof we must have $a_i = a_j = \frac{1}{\sqrt{r}}$. So all positive integers r and n. For

$$a_n^{(i)} = \frac{\sigma_{in}}{\sigma\sqrt{n}} = \frac{\sigma_{in}}{\sigma_{in}\sqrt{rn}} = a_i^{(n)} a_n^{(i)},$$

and we may let $n \to \infty$ to conclude that $a_i = a_j$.

7.4.3 *Examples.* It can be shown [see, for example, Breiman (1968, p. 204)] that X is stable if T is a measurable function such that $\mathbb{E}|\cos(tX+c)|^{\alpha}$ depends only on t, where α is one of the following two forms:

$$\alpha = \alpha_1 = \ldots \quad (1)$$

$$0 < \alpha < 1 \quad \text{or} \quad 1 < \alpha < 2 \quad \beta \neq 0, \quad \alpha = 1, \quad |\beta| < 1, \quad (2)$$

$$\alpha = \alpha_2 = \ldots \quad (3)$$

with α, β, σ as in (2). (Here $\alpha = 0$ when $\alpha = 0$.)

We shall prove only that a random variable with such a characteristic function must be stable. If $X_1, \ldots, X_n$ are iid with each X_i having characteristic function $\exp(-\sigma^{\alpha}|t|^{\alpha})$ with α of the form (2) or (3), then in case (2)

$$\mathbb{E}\exp(itX_1) \cdots \mathbb{E}\exp(itX_n) = \exp(-n\sigma^{\alpha}|t|^{\alpha}) = \mathbb{E}\exp(itn^{1/\alpha}X_1),$$

and in case (3)

$$\mathbb{E}\exp(itX_1) \cdots \mathbb{E}\exp(itX_n) = \exp\left(-n\sigma|t|\left[1 + i\beta\frac{2}{\pi}\operatorname{sgn}(t)\log|t| + i\beta\right]\right)$$

$$= \exp(-n\sigma|t|)\exp\left(-in\beta\frac{2}{\pi}\sigma t\log|t|\right).$$

This is of the form

$$|\phi(t)|^n = \exp(n\log|\phi(t)|) = \exp(n\log|\phi(t)|)\exp(in\theta(t)) = \phi(t)^n = \mathbb{E}\exp(itS_n)$$

where $\phi_0 = \phi(t) = e^{i\theta}$ in case (1) and $\phi_0 = \phi(t)e^{i\theta(t)}$ in case (2). Therefore $S_n \stackrel{d}{=} \sigma_0 X_1 + b_0$, with $\sigma_0 = n^{1/\alpha} = n^{1/\alpha}$ in each case. X is stable.

Now X has a symmetric distribution ($P[X \in B] = P[-X \in B]$ for all $B = (a,b)$) if T is symmetric [see 7.1.5(c)], so the general form of the symmetric stable characteristic function is

$$\phi(t) = \exp(-|\sigma t|^{\alpha}) \quad \alpha = 1, \quad 0 < \alpha < 2 \quad (4)$$

When $\alpha = 1$, one sees that $X = 0$ [primarily, if $\alpha = 2$ then $X = \sigma Z$ where Z is standard normal]. When $\alpha = 1$, X has the Cauchy density with parameter σ

If X is stable (not necessarily symmetric) and $0 < \alpha < 2$, then X is not differentiable at $t = 0$ so by (7.5.6), $E(|X|) = \infty$. If the symmetric case, $E(|X|)$ does not exist. For $\alpha = 2$, X is normal, $X \sim N(\mu, \sigma^2)$, and $E(|X|) = E(X^2) = \mu^2 + \sigma^2$, necessarily infinite. If $\alpha = 1$ and X has a Cauchy distribution, then $E(X^2) = \infty$, so that $\mathrm{Var}(X) = \infty$. This can be measured, for if X has finite mean and variance, the fact that X can be obtained as a limit of a sequence of arithmetic sums suggests that X must be normal (see the opening paragraphs of this section).

It can be shown (see Feller [1966, p. 215]) that if X is stable, $E(|X|^r) < \infty$ for all rth moments for $r < \alpha$ (or $r \le \alpha$).

Problems

1. The Lévy–Khintchine problem shows that the function $\lambda(t) = \exp\{-\sigma^\alpha |t|^\alpha [1 - i\beta \tan(\pi\alpha/2)\operatorname{sgn} t] + i\mu t\}$, $\sigma > 0$, $|\beta| \le 1$, $0 < \alpha \le 2$, and μ real, is a characteristic function (see 7.4.3). Let $X_1, \ldots, X_n$ be independent random variables, each uniformly distributed between $-\pi$ and $+\pi$. Define

$$Y_n = c_n \sum_{j=1}^{n} \frac{\sin jX_j}{j^{1/\alpha}}, \qquad c_n = \left(\sum_{j=1}^{n} \frac{1}{j^2}\right)^{-1/2}.$$

   If the X_j are the positions of masses distributed at random on $[-\pi, \pi]$, then Y_n is the gravitational force of the origin, assuming an inverse αth power law.

   (a) Show that the characteristic function of Y_n is

$$\lambda_n(t) = \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \prod_{j=1}^{n} \left[1 - \frac{|\sin(c_n t \sin jx)|^\alpha}{2\pi j}\right] dx_1 \cdots dx_n.$$

   (b) Show that $\lambda_n(t) \to \lambda(t) = \exp\{-\int_{-\pi}^{\pi} |\sin(c_n t \sin x)|^\alpha \, dx\}$ as $n \to \infty$.

   (c) Hint: To change α variable, use $\sin x = \sin^2(x/2) + \cos^2(x/2)$ to show that $\lambda(t)$ is of the form $\exp\{-\sigma(t)^\alpha\}$, $\sigma(t) > 0$, $|t| > 0$, $\sigma(0) = 0$. (The case $\alpha = 2$ corresponds to a normal distribution, and $\alpha = 0$ to a degenerate distribution, so these characteristic functions are universally represented.)

7.5 Infinitely Divisible Distributions

There are many laws that do not fit into any of the categories we have considered so far. For example, let T_n be the number of successes in n Bernoulli trials, with probability p_n of success on a given trial. Then T_n has the binomial distribution

$$P\{T_n = r\} = \binom{n}{r} p_n^r (1 - p_n)^{n-r}, \qquad r = 0, 1, \ldots, n.$$

If we set $r = 2, p_1 = p, p_2 = q$, we obtain $\mu_2 = 2$, then

$$P[T_2 = k] = \frac{k+1}{2^{k+2}}, \qquad k = 0, 1, \ldots,$$

(see Feller, 1950, p. 95, for details). A discrete random variable X with $P[X = k] = \alpha^{k+1}(k+1)$, $k = 0, 1, \ldots$, is said to have the *Pólya distribution*. In this case,

$$F_{T_2}(k) = P[T_2 \le k] = \sum_{j=0}^{k} P[T_2 = j] = \sum_{j=0}^{k} P[Y = j]$$
$$= P[X \le k] = F_X(k),$$

hence $T_2 \overset{d}{=} X$.

Now this can be regarded as a linear law for sums of independent random variables. We may represent T_2 as $X_{11} + X_{12} + \cdots + X_{1\nu}$, where X_{1i}, the number of successes on trial 1 in experiment T_2, the indicator of the event (success on trial 1), is 1 with probability p_1 and 0 with probability $1 - p_1$, and the Y_{1i} are independent. The difference between this case and the previous ones is that we are no longer dealing with a single sequence of random variables Y_1, a row of a single X array $X_{11}, X_{12}, \ldots$ are independent. Instead, for each ν we have a different sequence $X_{\nu 1}, \ldots, X_{\nu\nu}$.

We may summarize a more general version of the scheme as well as previous results, as follows. Construct a triangular array,

$$\begin{bmatrix} X_{11} & & & & \\ X_{21} & X_{22} & & & \\ X_{31} & X_{32} & X_{33} & & \\ \vdots & & & & \\ X_{\nu 1} & X_{\nu 2} & X_{\nu 3} & \cdots & X_{\nu\nu} \\ \vdots & & & & \end{bmatrix}.$$

We assume that for each ν, $X_{\nu 1}, \ldots, X_{\nu\nu}$ are independent. We say nothing as yet about any relation between rows. We set $T_\nu = X_{\nu 1} + \cdots + X_{\nu\nu}$; we want to investigate convergence in distribution of the sequence $\{T_\nu\}$.

Notice that if we are interested in expression $\lim_{\nu\to\infty} P[S_\nu = k]$, where S_ν is the sum of independent random variables $X_1, \ldots, X_\nu$, we may construct an appropriate triangular array by

$$X_{\nu 1} = \frac{X_1 - a_\nu}{b_\nu}, \quad \ldots,$$

etc.

$$T_\nu = \sum_{i=1}^{\nu} X_{\nu i} = \frac{S_\nu - a_\nu}{b_\nu}.$$

Thus the triangular array scheme includes the previous cases we have considered.

Note also that the Lindeberg theorem refers to the triangular array. If the X_{nj} have finite mean m_{nj} and finite variance σ_{nj}^2,

$$s_n^2 = \operatorname{Var}\left[\sum_{j=1}^{n} X_{nj}\right] = \sum_{j=1}^{n} \sigma_{nj}^2,$$

and for every $\varepsilon > 0$,

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|x - m_{nj}| \geq \varepsilon s_n} (x - m_{nj})^2 \, dF_{nj}(x) \to 0 \qquad \text{as} \qquad n \to \infty,$$

then

$$\frac{1}{s_n} \sum_{j=1}^{n} (X_{nj} - m_{nj}) \xrightarrow{\mathcal{L}} N(0,1) \qquad \text{normal } (0,1).$$

The proof is the same as in 7.5.1, with the distribution function F_j replaced by F_{nj}.

A natural question is the characterization of the possible limiting distributions of a triangular array. This problem was solved for normed sums in Section 7.6. However, as it stands, the question is not sensible, even if we assume that the triangular array comes from a single sequence of random variables. Let X be an arbitrary random variable, and take $X_{n1} = X$, $X_{nj} = 0$ for $n \geq 2$, $2 \leq j \leq n$. Then $S_n = \sum_{j=1}^{n} X_{nj} = X$, so no limit distribution is present. Thus some restriction must be imposed.

One way to take care of this difficulty is to assume the hypothesis of uniform asymptotic negligibility, as we did in considering the converse of the Lindeberg theorem.

$$\max_{1 \leq j \leq n} P[|X_{nj} - m_{nj}| \geq \varepsilon] \to 0 \qquad \text{as} \qquad n \to \infty \qquad \text{for every} \qquad \varepsilon > 0.$$

However, we will sacrifice generality for simplicity and assume that for each n, $X_{n1}, X_{n2}, \ldots, X_{nn}$ are identically distributed. We may now characterize the possible limiting distributions.

7.8.2 *Definition.* A random variable X (or its distribution function F) is *infinitely divisible* if for each n, X has the same distribution as the sum of n independent, identically distributed random variables. That is, for each n, we may write $X = X_{n1} + \cdots + X_{nn}$, where F_{nj} is the common distribution function of $X_{n1}, \ldots, X_{nn}$.

7.5.2 *Theorem.* The random variable X is infinitely divisible iff there is a triangular array $\{X_{n,k}\}$ with $X_{n,1}, \ldots, X_{n,k_n}$ iid for each n, such that $T_n = \sum_{k=1}^{k_n} X_{n,k} \xrightarrow{\mathcal{L}} X$.

Proof. Let X be infinitely divisible. For each n we may write $X = Y_{n,1} + \cdots + Y_{n,n}$ where the $Y_{n,k}$ are iid. Then $T_n = \sum_{k=1}^{n} Y_{n,k} = X$; hence $T_n \xrightarrow{\mathcal{L}} X$.

The converse is another application of Khinchin's weak convergence theorem. Assume we have a triangular array with $T_n \xrightarrow{\mathcal{L}} X$. Fix the positive integer r and let

$$T_n = Z_n^{(r)} + \cdots + Z_n^{(r)},$$

where

$$\begin{aligned} Z_n^{(r)} &= X_{n,1} + \cdots + X_{n,k_1}, \\ Z_n^{(r)} &= X_{n,k_1+1} + \cdots + X_{n,k_2}, \\ &\vdots \\ Z_n^{(r)} &= X_{n,k_{r-1}+1} + \cdots + X_{n,k_n}. \end{aligned}$$

Since $T_n \xrightarrow{\mathcal{L}} X$, we see from Theorem 7.2.3 that $\{T_n, n = 1, 2, \ldots\}$ is relatively compact. This means that the sequence of distribution functions is relatively compact. By 7.2.3, $\{Z_n^{(r)}, n = 1, 2, \ldots\}$ is tight. But

$$\begin{aligned} P(Z_n^{(r)} \le x) &= P(Z_n^{(r)} \le x, \ldots, Z_n^{(r)} \le x) \\ &\qquad \text{by independence of the } Z_n^{(r)} \\ &= P(T_n \le x)^r \end{aligned}$$

and similarly,

$$P(Z_n^{(r)} < -x) = P(T_n < -x)^r.$$

It follows that $\{Z_n^{(r)}, n = 1, 2, \ldots\}$ is tight, and hence relatively compact by 7.2.3. Thus we have a subsequence $\{Z_{n'}^{(r)}, n' = n_1', n_2', \ldots\}$ converging in distribution to a random variable Y_r for the $Z_{n'}^{(r)}$. By 7.2.5, $T_{n'} \xrightarrow{\mathcal{L}} Y_1 + \cdots + Y_r$, where $Y_1, \ldots, Y_r$ are iid with $Y_1 \stackrel{\mathcal{L}}{=} Y_r$. But $T_{n'} \xrightarrow{\mathcal{L}} X$; hence $X \stackrel{\mathcal{L}}{=} Y_1 + \cdots + Y_r$.

It can be shown (Gnedenko and Kolmogorov, 1954) that Theorem 7.5.1 still holds if the condition that the X_n are identically distributed is replaced by the condition

$$\lim_{n\to\infty} \frac{\max_{1\le k\le n} \sigma_k^2}{\sum_{k=1}^{n} \sigma_k^2} = 0.$$

7.5.2 *Examples of Infinitely Divisible Random Variables* (a) Every stable random variable is infinitely divisible. This may be seen from the fact that every stable X is a limit in distribution of a sequence of normalized sums, hence a limit of row sums of a triangular array in which the $Y_{nk} = a_n X_k + b_n$, $k = 1, \ldots, n$, have the same distribution. Alternatively, if $Y_1 = \cdots = Y_n \stackrel{\mathcal{L}}{=} a_n X + b_n$, then

$$X \stackrel{\mathcal{L}}{=} \sum_{k=1}^{n} \left[\frac{a_k}{a_n} X_k + \frac{b_k - b_n}{a_n} \right].$$

(b) A random variable X has Poisson type if it is infinitely divisible. Let F have the Poisson distribution $P(\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \ldots$. The characteristic function of F is

$$\phi(t) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = \exp\{\lambda(e^{it} - 1)\}$$

and it follows that if $X_1, \ldots, X_n$ are independent with X_k Poisson with parameter λ_k, $k = 1, \ldots, n$, then $X_1 + \cdots + X_n$ is Poisson with parameter $\lambda_1 + \cdots + \lambda_n$. In particular, if F is Poisson with parameter λ, then $F = F_n * \cdots * F_n$, where the F_n are Poisson with parameter λ/n. Thus F is infinitely divisible. Since all F is infinitely divisible, so is any F of the similar statement holds for shifted random variables). Hence a random variable of the Poisson type with $F = b + F$ Poisson, $\alpha \ge 0$, is infinitely divisible. The characteristic function of $aF = b$ is $\exp\{i\alpha t + \lambda(e^{i\alpha t} - 1)\}$.

(c) A random variable with the gamma distribution is infinitely divisible. For X has density

$$f(x) = \begin{cases} \dfrac{\alpha^{\beta} x^{\beta-1} e^{-\alpha x}}{\Gamma(\beta)} & x \ge 0, \\ 0 & x < 0 \end{cases}$$

where $\alpha, \beta > 0$. The characteristic function of X is

$$\phi(t) = (1 - it/\alpha)^{-\beta} = [1 - it/\alpha]^{-\beta}.$$

[illegible] X is the sum of n independent gamma distributed random variables with parameters α/n and β.

[illegible] some results on gamma [illegible] commonly encountered distributions

7.5.4 Theorem. If A_1 and A_2 are infinitely divisible characteristic functions then so is A_1A_2. If A is infinitely divisible, then $\bar{A}$, the complex conjugate of A, and $|A|^2$ are infinitely divisible.

Proof. If $A_j = (A_{jn})^n$, $j = 1, 2$, then $A_1A_2 = (A_{1n}A_{2n})^n$; since $A_{1n}A_{2n}$ is the characteristic function of the sum of two independent random variables with characteristic functions A_{1n} and A_{2n}. The last assertion is proved. If X has characteristic function A, then $-X$ has characteristic function $\bar{A}$ (see 7.1.5(c)); thus $\bar{A} = (\bar{A}_n)^n$, then $\bar{A} = (\bar{A}_n)^n$, hence $\bar{A}$ is infinitely divisible. Since $|A|^2 = A\bar{A}$, $|A|^2$ is also infinitely divisible. □

If $\mathcal{A}$ is the entire class of characteristic functions of random variables, the proof of 7.5.4 shows that if $A_1, A_2 \in \mathcal{A}$, then $A_1A_2 \in \mathcal{A}$.

Also if $A \in \mathcal{A}$, then $\bar{A} = A$ and $|A|^2 \in \mathcal{A}$. Furthermore, 7.5.8 implies that if $A_n \in \mathcal{A}$, $n = 1, 2, \ldots$, and $\lim_{n\to\infty} A_n(t) = A(t)$ for all t, where A is a characteristic function, then $A \in \mathcal{A}$. A similar result holds for infinitely divisible characteristic functions.

7.5.5 Theorem. If A_n is an infinitely divisible characteristic function for each $n = 1, 2, \ldots$, and $\lim_{n\to\infty} A_n(t) = A(t)$ for all t, where A is a characteristic function, then A is infinitely divisible.

Proof. Let X_n be a random variable with characteristic function A_n, $n = 1, 2, \ldots$. If r is a fixed positive integer, then $X_n = X_n^{(1)} + \cdots + X_n^{(r)}$, where the $X_n^{(k)}$, $k = 1, \ldots, r$, are [illegible]. If X is a random variable with characteristic function A, then $X_n \xrightarrow{\mathcal{L}} X$ by 7.1.9 so that $\{X_n\}$ is [illegible], and hence tight, by 7.3.1. Just as in the proof of 7.5.2, it follows that $\{X_n^{(k)}\}$ is tight. By 7.3.4 we have a subsequence $\{X_{n'}^{(1)}\}$, $n' = n_1', n_2', \ldots$, converging in distribution to a random variable Y_1; hence again as in 7.5.2, $X_{n'} - X_{n'}^{(1)} \to Y_2 + \cdots + Y_r$, where $Y_2, \ldots, Y_r$ are [illegible] with $X = Y_1 + \cdots + Y_r$. Thus $X_{n'} - X_{n'}^{(1)} \xrightarrow{\mathcal{L}} Z = Y_2 + \cdots + Y_r$. In other words, A is infinitely divisible. □

Now if A is infinitely divisible, a uniqueness question arises: namely, can it be represented in two different ways as the nth power of a characteristic

function! This is actually an exercise in complex variables, as follows. Let f and g be continuous complex-valued functions on the unbounded set S, with $f^n = g^n$, assume that $f(x) = g(y)$ for at least one $x \in S$. The set $S - \{x\}$ and $f - g^k = h$ since f and g are characteristic functions of random variables. $f(0) = g(0) = 1$ [...] f and g are never 0 on S, then $f = g$. Note that $f(x)^n = 1$, and therefore $f(x)$ is a continuous map of a unit [illegible] $e^{2\pi i/n}, \ldots, e^{2\pi i(n-1)/n}$ [...]. Since the image of S under $f(x)$ is connected, it must consist of a single point, thus $f(x)$ is a constant necessarily 1 because f and g agree at the origin. Thus the representation of h as the nth power of a characteristic function is unique, provided we can establish that an infinitely divisible characteristic function never vanishes.

7.5.6 *Theorem* If ϕ is an infinitely divisible characteristic function, then ϕ is never 0.

Proof If $\phi = |\phi|^{i\theta}$, then $|\phi|^2 = \phi\bar{\phi}$. Since $|\phi|^2$ is infinitely divisible by 7.5.4, we may as well assume that ϕ itself is real, and we need only consider $\phi > 0$. Thus $\phi_n = \phi^{1/n} = \exp\{(1/n)\log|\phi|\}$, and $\lim \phi(x) = 0$, then $\phi_n(x) \to 1$, and if $\lim \phi(x) = 0$, then $\phi_n(x) \to 0$ for all x. For any $\delta > 0$, $\lim \phi_n(x) = 1$ uniformly on some neighborhood of the origin. Thus ϕ_n converges to a function ψ that is 1 in a neighborhood of the origin. By 7.2.8, ψ is a characteristic function and hence continuous everywhere. But ψ takes on only the values 0 and 1, and hence $\psi \equiv 1$. Thus for any x fixed, $\phi_n(x) \to 1$, so that $\phi(x) \neq 0$ for sufficiently large n. Therefore $\phi(x) = [\phi_n(x)]^n \neq 0$. ∎

Example 7.5 Why is [illegible] the sense that random variables with characteristic functions can be used as building blocks for arbitrary infinitely divisible random variables.

7.5.7 *Theorem* The random variable X is infinitely divisible iff there is a sequence of sums $S_n = \sum_{k=1}^{k_n} X_{nk} \xrightarrow{\mathcal{L}} X$, where for each n the X_{nk} are independent (not necessarily identically distributed) and each X_{nk} is of the discrete type.

Proof Use "if" part follows from 7.5.6(b), the first condition in 7.5.1, and 7.5.5, so assume X infinitely divisible. Since ϕ, the characteristic function of X, is continuous and never 0, by 7.5.6, ϕ has a continuous logarithm, to be denoted by $\log \phi$ (more specifically, $\log \phi(0) = \log 1 = 0$, the logarithm is determined uniquely [see Ash 1967, p. 98] for a discussion of continuous logarithms). If $\lambda = (0, \infty)$, where ψ_λ is a characteristic function, then

$$\psi_\lambda = \left| \exp\left[\int_0^1 \log \lambda \right] \right|,$$

Hence as in the discussion before Theorem 7.6.2,

$$\phi_n = \exp\left\{\frac{1}{n}\log\phi\right\},$$

so for any fixed u we have

$$\begin{aligned}\exp\{\lambda_n(u)\} &= \phi\left\{\exp\left[\frac{1}{n}\log\phi(u)\right]\right\}^n \\ &= \left\{\frac{1}{n}\log\phi(u) + o\left(\frac{1}{n}\right)\right\}^n \qquad \text{since} \qquad e^x = 1 + x + o(x) \\ &\to \log\phi(u) \qquad \text{as} \qquad n \to \infty.\end{aligned}$$

Thus

$$\log\phi(u) = \lim_{n\to\infty} n\{\phi_n(u) - 1\} = \lim_{n\to\infty} n\int_{-\infty}^{\infty}(e^{iux} - 1)\,dF_n(x)$$

where F_n is the distribution function corresponding to ϕ_n. It follows from the dominated convergence theorem that for each A we can select a positive number $a = a(A)$ such that if $n > a$ and $u \le A$ then

$$\int_{-\infty}^{-a}(e^{iux} - 1)\,dF_n(x) + \int_{a}^{\infty}(e^{iux} - 1)\,dF_n(x) \le \frac{1}{A^2} \qquad \text{for all} \qquad u,$$

and we may then choose a positive integer $n = n(A)$ such that

$$\left|\int_{-a}^{a}(e^{iux} - 1)\,dF_n(x) - \sum_{j=1}^{k}(e^{iux_j} - 1)\{F_n(x_{j+1}) - F_n(x_j)\}\right| < \frac{1}{A^2}$$

for all $u \le A$, where $x_j = -a + 2aj/k$, $j = 0, \ldots, k$.

It follows that we may obtain $\log\phi$ as a pointwise limit of sums of the form

$$\sum_{j=1}^{k_n}\{\exp(iu_{nj}) - 1\}\lambda_{nj},$$

where $\lambda_{nj} = n\{F_n(x_j) - F_n(x_{j-1})\}$ and $u_{nj} = x_j$.

We conclude our discussion by mentioning the *Lévy–Khinchine representation*. The characteristic function ϕ is infinitely divisible iff

$$\log\phi(u) = iau - \frac{\sigma^2u^2}{2} + \int_{-\infty}^{\infty}\left\{e^{iux} - 1 - \frac{iux}{1+x^2}\right\}\frac{1+x^2}{x^2}\,dG(x),$$

where a is real, $\sigma^2 \ge 0$, and G is a finite measure on $\mathcal{B}(\mathbb{R})$ such that $G\{0\} = 0$.

This result is basic for a deeper study of the central limit theorem, in particular the sufficient conditions for convergence to a particular infinitely divisible distribution mentioned in Theorem 7.3.2. The sufficient conditions for convergence in Section 7.5 that are given by Gnedenko and Kolmogorov (1954) in Theorem 7.5.1 and the Lévy–Khintchine representation are also given by Cheng (1968) and Parker (1967).

Problems

1. The random variable X is said to have the geometric distribution if $P(X = k) = q^k - q^{k+1}$, $k = 0, 1, 2, \ldots$, where $0 < q < 1$ and $p = 1 - q$. Show that the characteristic function is given by

$$\phi(t) = p^2 \sum_{k=0}^{\infty} (qe^{it})^k = \frac{p^2}{1 - qe^{it}}.$$

   Is it infinitely divisible (see 7.3.7)?

2. Let $g(z) = \sum_{n=1}^{\infty} n^{-z}$, the zeta function of Riemann (see Titchmarsh, 1939). The series converges absolutely for $\mathrm{Re}\, z \geq 1 + \varepsilon$, any $\varepsilon > 0$. Also

$$g(z) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k^{-z}},$$

   where p_k is the kth prime (see Ash, 1971, Chapter 6 for details). If a is a real and random quantity, show that $\phi(t) = g(it)/g(ia)$ is an infinitely divisible characteristic function (see 7.3.7).

3. Give an example of an infinitely divisible characteristic function that is not stable.

4. When characteristic functions are not easy to manipulate, the following technique is sometimes useful for identifying the distribution of a sum of independent random variables.

   Let X and Y be independent random variables, and let $Z = X + Y$. If X, Y, and Z have distribution functions F_1, F_2, and F_3, show that F_3 is the convolution of F_1 and F_2; in symbols, $F_3 = F_1 * F_2$, that is,

$$F_3(z) = \int_{-\infty}^{\infty} F_1(z - y)\, dF_2(y).$$

   Provide the generalization of $F_1 * F_2$, that is,

$$F_1 * F_2(z) = \int_{-\infty}^{\infty} F_1(z - y)\, dF_2(y).$$

If X [respectively, Y] has density f_X [respectively, f_Y], show that Z has density f_Z, where

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx.$$

[If both X and Y have continuous, positive densities given by $\exp(-x^2)$ and $\exp(-y^2)$, respectively.]

Intuitively, the probability that X falls in $(x, x+dx)$ is $dF_X(x) = f_X(x)\,dx$ given that $X = x$, we have $Z = x + Y$. If $Y = y$, this happens with probability $F_Y(z-x)$. Integrate over x to obtain the total probability that $Z \le z$, namely, $F_Z = F_X * F_Y$. The other formulas have a similar interpretation.

Note also that convolution is associative, that is, $F_1 * (F_2 * F_3) = (F_1 * F_2) * F_3$. There is a symmetric way to prove this, but a probabilistic interpretation makes it unnecessary. For if X_1, X_2, and X_3 are independent random variables with distribution functions F_1, F_2, and F_3, respectively, then $(F_1 * F_2) * F_3$ is the distribution function of $(X_1 + X_2) + X_3$, while $F_1 * (F_2 * F_3)$ is the distribution function of $X_1 + (X_2 + X_3)$.

Finally, we note that if F is a distribution function of a random variable that is infinitely divisible, then for each n there is a distribution function F_n [of a random variable Y such that $Y = X_1 + X_2 + \cdots + X_n$ in distribution].

5. If $\varphi = \exp\psi$ and ψ is the characteristic function of a random variable, show that $\exp[\psi(t) - 1]$ is an infinitely divisible characteristic function.

7.6 Random Convergence to the Normal Law

Let $X_1, X_2, \ldots$ be independent, random variables with finite mean and variance, and suppose that $T_n = (S_n - nm)/\sqrt{n} \xrightarrow{\mathcal{L}} N(0, \sigma^2)$ [converges in law, very often the statement is made that "for large n, S_n is approximately normal with mean $nm = E(S_n)$ and variance $n\sigma^2$," that is,

$$P\{a < S_n < b\} \approx \int_{a}^{b} \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\left[-\frac{(x-nm)^2}{2n\sigma^2}\right] dx \quad \text{as} \quad n \to \infty.$$

Let us try to prove this. If X is normal (m, σ^2) and X^* is normal $(0, 1)$, then

$$|P\{S_n \le x\} - P\{X \le x\}| = \left|P\left\{T_n \le \frac{x - nm}{\sigma\sqrt{n}}\right\} - P\left\{X^* \le \frac{x - nm}{\sigma\sqrt{n}}\right\}\right|$$

$$\left|\Phi_n\left(\frac{x - nm}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{x - nm}{\sigma\sqrt{n}}\right)\right|$$

and this will approach 0 as $n \to \infty$ if $A - \varepsilon < F^*$ on $[a, b]$. In fact this does happen, the proof relying on the following two results.

7.6.2 *Theorem.* Let $F, F_1, F_2, \ldots$ be nondecreasing distribution functions on $\mathbb{R}$. Suppose that S contains all the discontinuity points of F.

$$\begin{aligned} F_n(x) &\to F(x), \\ F_n(-\infty) &\to F(-\infty), \\ A_n(x) &\to A(x) \qquad \text{for } x \qquad \in S, \\ A_n(x) &\to A(x) \qquad \text{for all} \qquad x \in \mathbb{R}, \end{aligned}$$

then $A_n \to A$ *uniformly on* $\mathbb{R}$.

Proof. Let $\varepsilon > 0$ be given. We seek to construct a partition $-\infty = y_0 < y_1 < \cdots < y_{k-1} < y_k = \infty$ such that

$$\begin{aligned} y_i &\in S, \qquad 1 \le i \le k-1, \\ F(y_{i-1}) &\le F(y_i) + \frac{\varepsilon}{2}, \qquad 2 \le i \le k-1, \\ A(y_{i-1}) &\le A(y_i) + \frac{\varepsilon}{2}, \qquad 2 \le i \le k-1, \end{aligned}$$

where $y_0 = -\infty$.

We let $y_1 = -\infty$ and define $x_1 = \sup\{x : A(x) < A(y_1) + \varepsilon/2\}$. If $x_1 < \infty$, then $A(x_1) \le A(y_1) + \varepsilon/2$, because $A(x_1) = A(x_1+) \le \varepsilon/2$. Hence $F(x_1) \le F(x_1+) \le A(x_1) + \varepsilon/2 \le A(y_1) + \varepsilon$. Also $F(x_1) > F(y_1) + \varepsilon/2$, for if not, $F(x_1+) \le F(y_1) + \varepsilon/2$ for some $x > x_1$, contradicting the definition of x_1.

Now if $F(x_1) < A(y_1)$, then $x_1 \in S$ by hypothesis, and we set $y_2 = x_1$. If $F(x_1) = A(y_1)$, then since S is dense and F is right-continuous, we can find $x_2 \in S$ such that $x_1 < x_2$ and $F(x_1) < F(y_1) + \varepsilon/2$. Thus in either case we obtain $y_2 > x_1$ such that $A(y_2) \le A(y_1) + \varepsilon/2$ and $F(y_2) \le F(y_1) + \varepsilon/2$. Continue by defining $x_2 = \sup\{x : A(x) < A(y_2) + \varepsilon/2\}$ and proceed as above. Since $F(y_{i+1}) \ge F(y_i) + \varepsilon/2$ and F is bounded, the process will terminate in a finite number of steps and produce the desired partition.

For $n = 1, 2, \ldots$ say $y_1 < \cdots < y_{k+1}$. Then

$$\begin{aligned} F_n(x) - F(x) &\le F_n(y_{i+1}) - F(y_i) \\ &\le F(y_{i+1}) + \frac{\varepsilon}{2} - F(y_i) \\ &\le \varepsilon \qquad \text{for large } n, \text{ since } y_{i+1} \in S \text{ or } y_{i+1} = \infty \end{aligned}$$

Also,

$$\begin{aligned} P(x) &= P\{X_n(\omega) \le x \text{ for all sufficiently large } n\} \\ &\le P\{x_1 - \tfrac{1}{n} < X_n(\omega) \le x_1 + \tfrac{1}{n} \text{ for large } n \text{ since } x_1 \ne 0 \quad \text{or} \quad x_1 = -\infty\} \\ &= F(x_1) - F(x_1 - \tfrac{1}{n}) \\ &\to 0 \end{aligned}$$

This contradiction implies that $x_1 = 0$ and, in a similar way, it follows that $x_2 = \infty$. Hence $X_n \to F$ uniformly on $\mathbb{R}$. □

5.6.2 **Theorem.** *Let $X_1, X_2, \ldots$ be a sequence of distribution functions on $\mathbb{R}$. Assume that $X_n(-\infty) = 0$ for all n, $F(-\infty) = 0$, F is continuous everywhere, and X_n converges weakly to F. Then X_n converges to F uniformly on $\mathbb{R}$.*

Proof. In 5.6.1 take $S = \mathbb{R}$. Since X_n converges weakly to F and F is continuous everywhere, $X_n(x) \to F(x)$ and $X_n(x-) \to F(x-)$ for all $x \in \mathbb{R}$. Since $X_n(-\infty) \to F(-\infty)$ is given by assumption, it remains to show that $X_n(x) \to F(x)$ for all x. If $x \in \mathbb{R}$,

$$F(x-) = \liminf_n X_n(x-) \le \liminf_n X_n(x) \le \limsup_n X_n(x) \le F(x)$$

If $x = \infty$, then $F(x) = F(x-)$ by continuity of F; hence $X_n(x) \to F(x)$. □

Problems

1 (Glivenko–Cantelli Theorem) Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with common distribution function F. Let $S(F) = \{x : x = \infty, x = -\infty$ is the empirical distribution function of $X_1, \ldots, X_n$ based on ω, that is,

$$F_n(x, \omega) = \frac{1}{n} \begin{bmatrix} \text{the number of terms among } X_1(\omega), \ldots, X_n(\omega) \\ \text{that are } \le x \end{bmatrix}$$

$$= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i(\omega))$$

For example, if $n = 3$, $X_1(\omega) = 2$, $X_2(\omega) = \frac{1}{2}$, $X_3(\omega) = 7$, then

$$F_3(x, \omega) = \begin{cases} 0 & x < \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} \le x < 2 \\ \frac{2}{3} & 2 \le x < 7 \\ 1 & x \ge 7. \end{cases}$$

Intuitively, the factor $n^{-1/2}$ in $S_n - n\mu$ should approximate $\sigma(n)$. Show that there is a sequence $\{a_n\}$ of positive numbers such that $\sum_{n=1}^{\infty} f(a_n x)$ converges uniformly for $x \in [-1, 1]$.

7.7 The Relationship Between Vague and Other Convergence Theorems

In this section we look at various results involving the simplest kind of convergence in distribution and other types of convergence. In particular, the Skorokhod construction frequently allows a very effective substitution of almost everywhere convergence for the weaker convergence in distribution. We begin with a result about convergence in distribution of sums, products, and quotients.

7.7.1 *Slutsky's Theorem.* If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, where c is a constant, then

(a) $X_n + Y_n \xrightarrow{d} X + c$;

(b) $X_n Y_n \xrightarrow{d} cX$;

(c) $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}$ if $c \neq 0$.

Proof. (a) Let F_n be the distribution function of X_n, G_n the distribution function of Y_n, F the distribution function of X, and H the distribution function of $X + c$. If $x_0 - c$ and $x_0 + c$ are points of continuity of F and $\varepsilon > 0$, we can find $\delta > 0$ such that $F(x_0 + c + \delta) - F(x_0 - c - \delta) < \varepsilon$. Therefore

$$G_n(x_0) \leq F_n(x_0 + c + \delta) + P[Y_n - c < -\delta] \leq \varepsilon$$

for

$$F_n(x_0 - c - \delta) \leq G_n(x_0) + P[Y_n - c > \delta] \leq \varepsilon$$

Let x_0 be a continuity point of H, so that $x_0 - c$ is a continuity point of F. Choose $\varepsilon > 0$ so that $x_0 + c + \varepsilon$ and $x_0 - c - \varepsilon$ are continuity points of F as well. Then

$$F(x_0 - c - \varepsilon) \leq \liminf_{n \to \infty} G_n(x_0) \leq \limsup_{n \to \infty} G_n(x_0) \leq F(x_0 + c + \varepsilon).$$

Since ε can be arbitrarily small, it follows that $G_n(x_0) \to F(x_0 - c) = H(x_0)$.

(b) We use the same notation as in (a), with $X_n + Y_n$ replaced by $X_n Y_n$ and $x_0 + c$ replaced by cx_0. First assume that $c > 0$. Then we can take $\varepsilon > 0$

without loss of generality, assume $-\infty < \frac{1}{2} < \infty$. The idea is essentially the same as in (a).

If $X_n \downarrow \frac{1}{2}$ and $Y_n \downarrow \frac{1}{2}$, then $X_n \le x/y$ and $Y_n \le x/y$, so that $X_n \wedge Y_n \le \frac{x}{y}$. Similarly, if $X_n \uparrow \frac{1}{2}$ and $Y_n \uparrow \frac{1}{2}$, then $x - x \le X_n \wedge Y_n$ for $x \le x/2$. Therefore,

$$G_n(x) \le F_n\left(\frac{x}{2}, \frac{x}{2}\right) + P[X_n \ge x/2] + P[Y_n \ge x/2]$$

and

$$F_n\left(\frac{x}{2}, \frac{x}{2}\right) \le G_n(x) + P[X_n \le x - \varepsilon] + P[Y_n \le x - \varepsilon].$$

Let x be a continuity point of F, so that $x/2$ is a continuity point of F. Choose $\varepsilon > 0$ so that $x/2 \pm \varepsilon$ are continuity points of F as well. Then

$$F\left(\frac{x}{2} - \varepsilon\right) \le \liminf_{n \to \infty} G_n(x) \le \limsup_{n \to \infty} G_n(x) \le F\left(\frac{x}{2} + \varepsilon\right)$$

and since ε can be taken as small as we wish, it follows that $G_n(x) \to F(x/2) = F(x)$. [We have assumed that $Y_n > 0$ in the above manipulations, but this causes no difficulty because $Y_n \ge \varepsilon > 0$ except on a set of small probability. A similar comment applies in the proof of (c) below.]

If $c = 0$ we must show that $X_n Y_n \xrightarrow{p} 0$. This may be accomplished by observing that $\{|X_n Y_n| \ge \varepsilon\} \subset \{|X_n| \ge \varepsilon/k\} \cup \{|Y_n| \ge k\}$, and given $\varepsilon > 0$, we have $P[|X_n Y_n| \ge \varepsilon] \to 0$ as $n \to \infty$. Thus for sufficiently large n, we have $|X_n Y_n| < \varepsilon$ except on a set of small probability.

(c) As in (b) we can assume without loss of generality that $c > 0$. We use the same notations as in (a) with $X_n - Y_n$ replaced by X_n/Y_n and $X - Y$ replaced by X/Y. If $X_n/Y_n \le x$ and $|Y_n - c| < \varepsilon$, then $X_n \le x(c + \varepsilon)$ and $Y_n \ge c - \varepsilon$. Similarly, if $X_n \ge x(c - \varepsilon)$ and $|Y_n - c| < \varepsilon$, then $x < X_n/Y_n$ and $Y_n \le c + \varepsilon$. Therefore,

$$G_n(x) \le P[X_n \le x(c + \varepsilon)] + P[|Y_n - c| \ge \varepsilon]$$

and

$$F_n(x(c - \varepsilon)) \le G_n(x) + P[|Y_n - c| \ge \varepsilon].$$

The argument is completed as in (a), with $x(c \pm \varepsilon)$ replaced by $x(c - \varepsilon)$ by $x(c + \varepsilon)$ and $x(c \mp \varepsilon)$ by $x(c - \varepsilon)$. $\square$

We now turn to Skorokhod's construction.

5.4.2 *Skorokhod's Theorem.* Assume that $X_n \xrightarrow{\mathcal{L}} X$, and let F_n be the distribution function of X_n, with F the distribution function of X. Then there are random variables $Y_n (n = 1, 2, \ldots)$ and Y on $(\Omega, \mathcal{F}, P)$ with Borel sets and P-measure such that Y_n has the same distribution as X_n, Y has the same distribution as X, and $Y_n(\omega) \to Y(\omega)$ for every $\omega \in \Omega$.

Proof. The idea is easy since F_n converges pointwise to F except possibly at denumerable points of F, the inverse of F_n should converge to the inverse of F. More precisely, the distribution function F is increasing and right-hand continuous, so each $\omega \in (0, 1)$ will be a minimum value of x, say $x = x_0$, such that $F(x) \geq \omega$, and we have $F(x) \geq \omega$ for all $x \geq x_0$ but $F(x) < \omega$ for all $x < x_0$. We define

$$F^{-1}(\omega) = \min\{x : F(x) \geq \omega\}. \qquad \text{see Fig. 5.7...}$$

From the definition we have

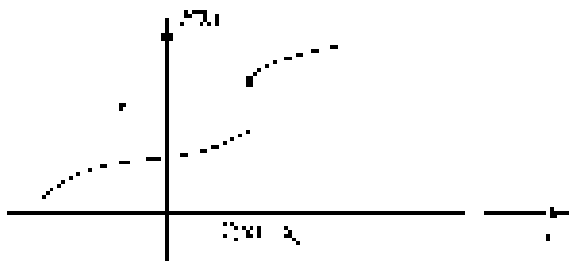

Figure 5.7...

(1) $F^{-1}(\omega)$ is finite, has upward height at all x of value at least $F(x) \geq \omega$. Similarly we define $F_n^{-1}(\omega) = \min\{x : F_n(x) \geq \omega\}$, then $F_n(F_n^{-1}(\omega)) \geq \omega$.

(2) $x < F^{-1}(\omega)$ implies $F(x) < \omega$; $x > F^{-1}(\omega)$ implies $F(x) \geq \omega$; and $F(F^{-1}(\omega) - \varepsilon) < \omega < F(F^{-1}(\omega) + \varepsilon)$ for arbitrary $\varepsilon > 0$. We will prove next

(3) $\liminf_{n \to \infty} F_n^{-1}(\omega) \geq F^{-1}(\omega)$.

For $\omega = 0$ or $\omega = 1$ (3) is obvious. Choose a continuity point $x < F^{-1}(\omega)$ such that $F(x) < \omega$ (see Fig. 5.8). By (1), $F_n(x) \to F(x)$ implies $F_n(x) < \omega$ for n sufficiently large, so $F_n^{-1}(\omega) \geq x$ for n sufficiently large. Thus $\liminf_{n \to \infty} F_n^{-1}(\omega) \geq x$, and since x is arbitrary, (3) must follow.

Figure 5.8

The analogous result for $\limsup_{n\to\infty} X_n$ fails because $\{\omega : F(\omega) \le x\}$ is not equivalent to $\{\omega : F(\omega) < x\}$. For example, see Fig. 7.7.1 with $x = x_0$.

(D) If F is continuous at x then $\limsup_{n\to\infty} F_n(x) \le F(x)$.

Let ω and ω' be in Ω. Let $\omega \neq \omega'$, and let $x > 0$. Choose a continuity point y of F such that $F(\omega) < y < F(\omega') + x$ (Fig. 7.7.3). Now $F(\omega') + x > y$ implies $\omega' \in \{\omega : F(\omega) < y\}$, and since $F(\omega)$ is the first point at which F reaches height ω, we have $F(\omega) \ge \omega$. Thus $\omega' \in \{\omega : F(\omega) < y\}$ (Fig. 7.7.4).

Figure 7.7.3

Figure 7.7.4

But $\omega \notin \{\omega : F(\omega) < x\}$ so not all sufficiently large n, $\omega \notin \{\omega : F_n(\omega) < y\}$ so $F_n(\omega) \ge y$ (Fig. 7.7.5). Consequently,

$$\limsup_{n\to\infty} F_n(\omega) \le F(\omega') + x \quad \text{for all } \omega' > \omega.$$

and the result follows.

By definition, $F(y)$ increases with y, and therefore has at most countably many discontinuities. Since a countable set has Lebesgue measure zero, we are free to change the definitions of F_n and F at points where F is discontinuous. In particular, the assumption $\mu\{\omega : Y(\omega) = y\} = 0$ for all y considered in the theorem now follows from (2) and (D). □

Part (a) of the next theorem gives a typical application of the Skorokhod construction (cf. 2.5 [2]).

7.7.3 Convergence of Transformed Sequences. Let $X, X_1, X_2, \ldots$ be random variables on $(\Omega, \mathcal{F}, P)$ and let $g : \mathbb{R} \to \mathbb{R}$ be continuous on $\{x : p(x) > 0\}$. Then

(a) $X_n \to X$ a.s. implies $g(X_n) \to g(X)$ a.s.

(b) $X_n \xrightarrow{P} X$ implies $g(X_n) \xrightarrow{P} g(X)$

(c) $X_n \xrightarrow{\mathcal{D}} X$ implies $g(X_n) \xrightarrow{\mathcal{D}} g(X)$

PROOF (a) Suppose that $X_n(\omega) \to X(\omega)$ for $\omega \in A$, where $P(A) = 1$, and ε is continuous on S, where $P_X(S) = 1$. Given $\delta > 0$, let $B = \{x : |x - y| < \delta \text{ for some } y \in S\}$. Then B is open and $S \subset B^\circ$. If $\omega \in A$ and $X(\omega) \in B$, then $X(\omega) \in S$, so $g(X_n(\omega)) \to g(X(\omega))$. Since $P(A \cap \{X \in B\}) = 1$, the result follows.

(b) If $g(X_n)$ fails to converge in probability to $g(X)$, then there exist $\varepsilon > 0$ and $\delta > 0$, and a subsequence of $1, 2, \ldots$, such that on the entire subsequence,

$$P\{|g(X_n) - g(X)| > \varepsilon\} > \delta.$$

We can then extract a further subsequence on which $X_n \to X$ a.s., and consequently $g(X_n) \to g(X)$ a.s. by (a). On the second subsequence we must have $P\{|g(X_n) - g(X)| > \varepsilon\} > \delta$, which contradicts the choice of the first subsequence.

(c) By Theorem 4, there are random variables $Y_n \stackrel{d}{=} X_n$ and $Y \stackrel{d}{=} X$ with $Y_n \to Y$ a.s., and so $g(Y_n) \to g(Y)$ a.s. By Theorem 3, $g(Y_n) \stackrel{d}{=} g(X_n)$. But $X_n \stackrel{d}{=} F_n$ implies $g(Y_n) \stackrel{d}{=} g(F_n)$, and $Y \stackrel{d}{=} F$ implies $g(X) \stackrel{d}{=} g(F)$. Thus $g(F_n) \xrightarrow{d} g(X)$. □

Problems

1. If $X_n \xrightarrow{d} X$, $a_n \to a$, and $b_n \to b$, show that $a_n X_n + b_n \xrightarrow{d} aX + b$.
2. If $X_n \xrightarrow{d} X$, show that $\{|X_n|^r : n \geq 1\}$ is uniformly integrable for $0 < r < \infty$.
3. If $Y_n \xrightarrow{d} X$ and the Y_n are uniformly integrable, show that X is integrable and $E(Y_n) \to E(X)$.

7.8 The *k*-Dimensional Central Limit Theorem

Just as the sum of a large number of independent random variables is, under mild conditions, approximately normal, the sum of a large number of independent random vectors has, approximately, a Gaussian (also called multivariate normal) distribution. In this section we give a version of the central limit theorem for independent random vectors.

7.8.1 Definitions and Comments First we review some definitions and results from 5.4 and 5.8. On $\mathbb{R}^k$ we have a partial ordering: $x \leq y$ iff $x_i \leq y_i$ for each $i = 1, \ldots, k$, and $x < y$ iff $x_i < y_i$ for every $i = 1, \ldots, k$.

If $F : \mathbb{R}^k \to \mathbb{R}$, we say that F is right-continuous at x iff $\lim F(y_n) = F(x)$ for every sequence $y_1 \geq y_2 \geq \cdots \to x$. If F satisfies the property that $a \leq b$ implies $F(a) \leq F(b)$, we may assume in the definition of right continuity that the y_n are merely required to

If $F:\mathbb{R}^1 \to \mathbb{R}^1$ we say that F is *continuous from the right* if $\lim_{t \to x+} F(t) = F(x)$ for every $x \in \mathbb{R}^1$. This is not equivalent to asymptotic continuity from the right (Exercise 2.5.1), but the two properties imply each other in general.

An increasing, right-continuous function $F:\mathbb{R}^1 \to \mathbb{R}^1$ is called a *distribution function* on $\mathbb{R}^1$. If F is a distribution function, then there is a Lebesgue–Stieltjes measure μ on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$ such that $\mu((a,b]) = F(b) - F(a)$ for $-\infty < a < b < \infty$. Conversely, if μ is a Lebesgue–Stieltjes measure on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$, there are several distribution functions F such that $\mu((a,b]) = F(b) - F(a)$ for all $-\infty < a < b < \infty$. If μ is a finite measure, it is usual to use the distribution function $G(t) = \mu((-\infty, t])$. That is the one we will use in most, though not all, applications between μ and G.

2.5.2 *Theorem.* Let μ be a finite measure on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$, and let $G(t) = \mu((-\infty, t])$. Then G is a distribution function such that $\lim_{t \to -\infty} G(t) = 0$. Conversely, suppose that G is a continuous function on $\mathbb{R}^1$ with the property that for every $\varepsilon > 0$ there is an $M > 0$ such that if $-\infty < a < b < \infty$ and $G(b) - G(a) > \varepsilon$, then $b - a < M$. Then there is a unique Lebesgue–Stieltjes measure μ on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$ such that $\mu((a,b]) = G(b) - G(a)$ for $-\infty < a < b < \infty$. We have $G(t) \geq G(0)$ when $t \geq 0$. The measure μ is finite if and only if $\sup\{G(t) : -\infty < t < \infty\} < \infty$.

Proof. The first part is left as an exercise following Example 2.5.1. If G is a distribution function and μ is a nonnegative measure such that $\mu((a,b]) = G(b) - G(a)$ for $-\infty < a < b < \infty$, then $\mu(\mathbb{R}^1) = \lim_{n \to \infty} \mu((-n, n]) = \lim_{n \to \infty} G(n) - G(-n)$. If $G(n) - G(-n) \leq \varepsilon$ for all sufficiently large n, then $\mu(\{x : |x| > n\}) \leq \varepsilon$ for all sufficiently large n, so that $G(t) \to 0$ as $t \to -\infty$.

Suppose now that G is a continuous function on $\mathbb{R}^1$ with the property that for every $\varepsilon > 0$ there is an $M > 0$ such that if $-\infty < a < b < \infty$ and $G(b) - G(a) > \varepsilon$, then $b - a < M$. Then there is a unique Lebesgue–Stieltjes measure μ on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$ such that $\mu((a,b]) = G(b) - G(a)$ for $-\infty < a < b < \infty$. If $-\infty < a < b < \infty$ and $G(b) - G(a) > \varepsilon$, then $\mu((a,b]) > \varepsilon$, so that $G(t) \to \infty$ as $t \to \infty$. If $-\infty < a < b < \infty$ and $G(b) - G(a) > \varepsilon$, then $\mu((a,b]) > \varepsilon$, so that $G(t) \to -\infty$ as $t \to -\infty$. Therefore $G(t) = \mu((-\infty, t])$. $\square$

In one dimension, G is continuous at x if $\mu(\{x\}) = 0$. We can formulate this condition in such a way that it generalizes to the higher-dimensional case, as follows.

Figure 9.8.1. Proof of Theorem 9.8.2.

PROOF. If G is continuous at x, then $\lim_{n\to\infty} G_n(x) = \lim_{n\to\infty} G_n(x-) = G(x)$ and $\lim_{n\to\infty} G_n(x+) = G(x+) = G(x)$. Conversely, if $\lim_{n\to\infty} G_n(x-) = \lim_{n\to\infty} G_n(x+)$, then

$$\lim_{n\to\infty} G_n(x) = \lim_{n\to\infty} G_n(x-) = \lim_{n\to\infty} G_n(x+) = G(x+)$$

by right-hand continuity of G.

Choose $\varepsilon > 0$, and choose points x and y such that $x < y$, $G(x) < \varepsilon/2$ and $G(y) > 1 - \varepsilon/2$ (see Fig. 9.8.1). If $n \in \{1, 2, \ldots\}$ we have $G_n(x) \le G(x) \le G(y) \le G_n(y)$, and therefore $G_n(x) \le G(x) < \varepsilon/2$ and $G(y) \le G_n(y) + \varepsilon/2$. Thus $\sup_{-\infty<x<\infty} |G_n(x) - G(x)| \le \varepsilon$.

We now consider the hyperplanes that are perpendicular to one of the coordinate axes, that is, sets of the form $\{x : x_j = a\}$, where $j \in \{1, \ldots, k\}$ and $a \in \mathbb{R}$. As the measure μ is finite, only countably many of these hyperplanes can have strictly positive μ-measure (see Sec. 1.2, Problem 12). If $\mathcal{H}_0$ is the set of these hyperplanes, we have the following result.

9.8.4 THEOREM. *If μ does not assign positive mass to any of the hyperplanes in $\mathcal{H}_0$, then Γ is continuous at μ.*

PROOF. This follows immediately from Corollary 9.8.3.

Recall that a sequence of finite measures μ_n on $\mathcal{B}^k$ converges weakly to a finite measure μ if $\int f\,d\mu_n \to \int f\,d\mu$ for every bounded continuous $f : \mathbb{R}^k \to \mathbb{R}$ (see Section 7.3). This gives a topological equivalent to the notion of weak convergence, and as the theorem was proved for finite measures on the Borel sets of an arbitrary metric space, it remains valid in $\mathcal{R}^k$. We may also establish a multidimensional analog of Theorem 9.5.6, as follows.

9.8.5 THEOREM. *Let $\mu_1, \mu_2, \ldots$ be finite measures on $(\mathbb{R}^k, \mathcal{B}^k)$ with corresponding distribution functions $F_1, F_2, \ldots$, where we take only*

$= \mu_n(-\infty, x]$ and $F(x) = \mu(-\infty, x]$. The following conditions are equivalent:

(a) $\mu_n \Rightarrow \mu$ weakly.

(b) $\mu_n(\mathbb{R}^1) \to \mu(\mathbb{R}^1)$ as $n \to \infty$ and $F_n(x) \to F(x)$ at each continuity point x of F in $\mathbb{R}^1$.

(c) $\mu_n(\mathbb{R}^1) = \mu(\mathbb{R}^1)$ for each n and $F_n(x) \to F(x)$ at each point $x \in \mathbb{R}^1$ that does not lie in any of the hyperplanes of $\mathcal{C}_0$.

Proof. (a) implies (b) by Theorem 7.8.1 above (take $f \equiv 1$) and we know that $\mu_n(\mathbb{R}^1) \to \mu(\mathbb{R}^1)$. If x is a continuity point of F, then $\mu\{x\} = \mu(-\infty, x] - \mu(-\infty, x) = F(x) - F(x-) = 0$ by Theorem 2.8.1(iv).

(b) implies (c) This follows from Theorem 7.8.1.

(c) implies (a) Given $\varepsilon > 0$, we can find a positive number T such that none of the vertices of $(-T, T]^k$ lies in a hyperplane of $\mathcal{C}_0$, and such that $\mu\{x : |x| = T\} < \varepsilon$. Then $\mu_n(-T, T]^k = \mu(-T, T]^k$ and $\mu_n(\mathbb{R}^k) - \mu(\mathbb{R}^k)$ by (c), and therefore $\mu_n(\mathbb{R}^k) - \mu_n(-T, T]^k) < 2\varepsilon$ for all sufficiently large n.

Let f be a bounded continuous function on $\mathbb{R}^k$. f is uniformly continuous on $[-T, T]^k$, so using finite chains and the $\mathcal{A}_0$ we can cut $[-T, T]^k$ into a finite number of rectangles $[r, s]$ such that if x and y belong to the same rectangle then $|f(x) - f(y)| < \varepsilon$. This enables us to approximate f on $[-T, T]^k$ by a function g such that g is constant on each $[r, s]$ and $|f - g| < \varepsilon$ on $(-T, T]^k$. Then, setting $M = \|f\| = \sup_x |f(x)|$, we have

$$\begin{aligned}
&\left| \int_{(-T,T]^k} f(x)\,d\mu_n - \int_{(-T,T]^k} f(x)\,d\mu \right| \\
&\quad \le \int_{(-T,T]^k} |f(x) - g(x)|\,d\mu_n + \left| \int_{(-T,T]^k} g(x)\,d\mu_n - \int_{(-T,T]^k} g(x)\,d\mu \right| \\
&\quad\quad + \int_{(-T,T]^k} |g(x) - f(x)|\,d\mu \\
&\quad \le \varepsilon \mu_n(-T, T]^k + \left| \int_{(-T,T]^k} g(x)\,d\mu_n - \int_{(-T,T]^k} g(x)\,d\mu \right| + \varepsilon \mu(-T, T]^k \\
&\quad \le \varepsilon + \left| \int_{(-T,T]^k} g(x)\,d\mu_n - \int_{(-T,T]^k} g(x)\,d\mu \right| + \varepsilon.
\end{aligned}$$

The right side of this inequality is less than some constant multiple of ε for sufficiently large n, and the result follows. ∎

We now begin the study of weak compactness on $\mathbb{R}^k$.

7.8.6 *Helly's Theorem* Let $\mu_1, \mu_2, \ldots$ be finite measures on $\mathbb{R}^k$ with corresponding distribution functions $F_1, F_2, \ldots$. As above we take $F_n(x) = \mu_n(-\infty, x]$. If $\mu_n(\mathbb{R}^k) \le M < \infty$ for all n, there is a distribution function

From $\mathbb{R}$ and a subsequence $\{F_{n_k}\}$ such that $F_{n_k}(x) \to F(x)$ at every continuity point x of F in $\mathbb{R}$.

Proof. The procedure is essentially the same as in 7.2.1. But we must be more careful sometimes in choosing the constants a_n. Let $\mathcal{A}$ be a dense subset of $\mathbb{R}$, and take $\mathcal{A} = \mathbb{Q}$. Then construct F and the sequence $\{n_k\}$ exactly as in 7.2.1. We must show that F is increasing. That is, $F(a) \le F(b)$ when $a \le b$. To accomplish this we must be a bit more careful of type of slightly larger than a and b. Then $a < c < b$ and $|c - a|$ and $|b - c|$ are small enough we have $F(a) \le F(c)$ and $F(c) \le F(b)$. But $F(a) = F(a^+)$ and $F(b) = F(b^+)$, and the inequalities follow. $F(x) \to 1$ as $x \to \infty$ and $F(x) \to 0$ as $x \to -\infty$ follow from the fact that F is right continuous and that $F_{n_k}(x) \to F(x)$ when x is a continuity point of F in $\mathbb{R}$. The sequence $\{a_{n_k}\}$ is bounded. For if $a_{n_k} \to \infty$ as $k \to \infty$ for some subsequence $\{n_k\}$, then $F_{n_k}(x) \to 1$ for all $x \in \mathbb{R}$. But $F(x) = 0$ for all $x \in \mathbb{R}$. Similarly, $\{a_{n_k}\}$ is bounded below. Finally, F is continuous from the right. For if $x \in \mathbb{R}$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $F(x + \delta) < F(x) + \varepsilon$. Then $F_{n_k}(x + \delta) < F(x) + \varepsilon$ for sufficiently large k. Hence $F(y) < F(x) + \varepsilon$ for all $y \in (x, x + \delta)$. Therefore $F(y) \to F(x)$ as $y \to x^+$. Now the right continuity of F implies that $F(x) = F(x^+)$ at least one x. Then there exists, by 7.2.1, there is a measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F(x) = \mu(-\infty, x]$ for all x and μ is finite because $\sup_x F(x) \le 1$. Lightness also implies that $\mu_{n_k}(\mathbb{R}) \to \mu(\mathbb{R})$ and by 7.2.5, μ_{n_k} converges weakly to μ. This proves relative compactness. ∎

7.2.8 *Remarks.* (a) Prokhorov's theorem is true in any complete, separable metric space (S, d). (b) $\mathcal{M}_1(S)$ is a Polish space.

(c) Prokhorov's theorem is true in any complete, separable metric space (S, d). (d) $\mathcal{M}_1(S)$ is a Polish space.

If the sequence of finite measures $\mu_1, \mu_2, \ldots$ is tight and there is a finite measure μ such that any weakly convergent subsequence of $\{\mu_n\}$ converges weakly to μ, then the entire sequence μ_n converges weakly to μ. For the Prohorov's theorem, see e.g. any weakly convergent subsequence

Characteristic functions can be defined in a natural way in $\mathbb{R}^k$

7.8.9 *Definition.* Let μ be a finite measure on $\mathcal{R}^k$. The characteristic function of μ is the mapping from $\mathbb{R}^k$ to $\mathbb{C}$ given by

$$\hat{\mu}(t) = \int_{\mathbb{R}^k} \exp(i\langle t, x\rangle)\, d\mu(x) \qquad t \in \mathbb{R}^k.$$

where $\langle t, x\rangle = \sum_{j=1}^{k} t_j x_j$.

If $\mu = P_X$, where X is a random vector $(X_1, \ldots, X_k)$, then $\hat{\mu}(t) = E\exp(i\langle t, X\rangle)$.

It follows that Theorem 7.1.2 can be extended to independent random vectors; in other words, the characteristic function of a sum of independent random vectors is the product of the individual characteristic functions.

As in one dimension, the characteristic function of a finite measure determines the measure uniquely, because of the following result.

7.8.10 *Inversion Formula.* If $\hat{\mu}$ is the characteristic function of a finite measure μ, and if $A = [a_1, b_1] \times \cdots \times [a_k, b_k]$ is a rectangle in $\mathbb{R}^k$ such that $\mu(\partial A) = 0$, then

$$\mu(A) = \lim_{c \to \infty} \frac{1}{(2\pi)^k} \int_{-c}^{c} \cdots \int_{-c}^{c} \prod_{j=1}^{k} \left[\frac{\exp(-ita_j) - \exp(-itb_j)}{it_j} \right] \hat{\mu}(t)\, dt$$

Proof.

$$\begin{aligned} \text{Let } I_c &= \frac{1}{(2\pi)^k} \int_{-c}^{c} \cdots \int_{-c}^{c} \prod_{j=1}^{k} \left| \frac{\exp(-ita_j) - \exp(-itb_j)}{it_j} \right| \hat{\mu}(t)\, dt \\ &= \frac{1}{(2\pi)^k} \int_{-c}^{c} \cdots \int_{-c}^{c} \prod_{j=1}^{k} \frac{\exp(-ita_j) - \exp(-itb_j)}{it_j} \\ &\quad \times \int_{\mathbb{R}^k} \exp\left[i \sum_{m=1}^{k} t_m x_m \right] d\mu(x)\, dt \\ &= \int_{\mathbb{R}^k} \cdots \int \prod_{m=1}^{k} f(a_m, b_m, x_m, c)\, d\mu(x) \end{aligned}$$

where $f(a, b, x, c) = \frac{1}{2\pi} \int_{-c}^{c} \left| \frac{\sin(t(x-a))}{t} - \frac{\sin(t(x-b))}{t} \right| dt$

As in 7.6.1 we have $\lim_{n\to\infty} f_n(x_n, y_n, t_n) = f(x, y, t)$, where $f \neq 0$ for $x \in \mathbb{R}$, $y, t > 0$. Therefore $\varphi_n(x_n, t_n) \to \frac{1}{2}$ for $x = -\infty$ or ∞. Therefore

$$\lim_{n\to\infty} I_n = \int_{\mathbb{R}^2} K(x, u)\, du\, dv,$$

where K is 0 on $\mathbb{R} \times [0, 1]$ and $\frac{1}{2}$ on $\mathbb{R} \times \{0, 1\}$. On the boundary of $A = [0, 1]^2$, K assumes values that are integral powers of $\frac{1}{2}$. It follows that $\lim_{n\to\infty} I_n = \mu(A)$ if $\mu(\partial A) = 0$. □

Here is the Kolmogorov–Prohorov version of Lévy's theorem 7.2.6.

7.6.22 *Theorem.* Let $\{\mu_n : n = 1, 2, \ldots\}$ be a sequence of measures on $\mathbb{R}^k$ such that $\sup_n \int |x|^2 \, d\mu_n < M < \infty$ for all n, and let φ_n be the characteristic function of μ_n. Let μ be a finite measure on $\mathbb{R}^k$ with characteristic function φ. Then $\mu_n \to \mu$ weakly iff $\varphi_n(t) \to \varphi(t)$ for every $t \in \mathbb{R}^k$.

Lemma. If $\mu_n \xrightarrow{v} \mu$ then $\varphi_n(t) \to \varphi(t)$ for $t \in \mathbb{R}^k$, uniformly on compact subsets of $\mathbb{R}^k$. We consider vectors $x = (x_1, \ldots, x_k)$ as rows and the coordinate vector e_j has a 1 in position j and 0 elsewhere. Then

$$\hat{\mu}^{(j)}(y) = \varphi_\mu(te_j) = \int_{\mathbb{R}^k} e^{ity_j} \, d\mu(x)$$

is in fact the Laurent transform of the measure $\mu^{(j)}$ on $\mathbb{R}$ defined by

$$\mu^{(j)}[a, b] = \mu_k(F_j^{-1}[a, b]) \qquad a < b \in \mathbb{R}$$

where F_j is the projection $(x_1, \ldots, x_k) \to x_j$ on the jth coordinate axis. (See 1.6.12 with $T = F_j$, $\nu = \mu$ and $\mathcal{E}' = \mathbb{R}^k$.)

Since $\varphi_n^{(j)}(y) = \varphi_{\mu_n}(te_j) = \varphi_n(te_j)$ as $t \to \infty$, it follows from 7.6.8 that the $\mu_n^{(j)}$, $n \geq 1$, are tight measures on $\mathbb{R}$, so we can find a positive number c_j such that $\mu_n^{(j)}[-c_j, c_j] > \frac{1}{2}$ for all n. Then $\mu_n[-c, c]^k > \prod_{j=1}^k [-c_j, c_j] > \frac{1}{2}$. Consequently, the μ_n, $n \geq 1$, are tight measures on $\mathbb{R}^k$. The condition $\sup_n \int |x|^2 \, d\mu_n < \infty$ assures that every weakly convergent subsequence of $\{\mu_n\}$ converges weakly to μ. By 7.6.8, $\mu_n \xrightarrow{w} \mu$. □

The following result, which characterizes weak convergence of k-dimensional probability measures in terms of weak convergence on $\mathbb{R}$, is a key step in the proof of the multivariate central limit theorem.

7.8.22 *Cramér-Wold Device.* Let $\mathbf{X}_n = (X_{n1}, \ldots, X_{nk})$, $n \geq 1$, be a sequence of k-dimensional random vectors. Then for $\mathbf{X}_n$ converge weakly to the random vector $\mathbf{Y} = (Y_1, \ldots, Y_k)$ it is necessary and sufficient that

$$\sum_{j=1}^{k} a_j X_{nj} \xrightarrow{d} \sum_{j=1}^{k} a_j Y_j \quad \text{for every} \quad \mathbf{a} = (a_1, \ldots, a_k) \text{ in } \mathbb{R}^k. \tag{1}$$

Proof. Let ϕ_n be the characteristic function of $\mathbf{X}_n$ and ϕ the characteristic function of $\mathbf{Y}$. Then by 7.8.9, condition (1) is equivalent to

$$\phi_n(a_1, \ldots, a_k) \to \phi(a_1, \ldots, a_k) \quad \text{for all} \quad n \to \infty \quad \text{and} \quad \mathbf{a} \in \mathbb{R}^k. \tag{2}$$

But if (2) holds, then taking $n = 1$ and applying 7.8.1 we have $\mathbf{X}_n \xrightarrow{d} \mathbf{Y}$. Conversely, if $\mathbf{X}_n \xrightarrow{d} \mathbf{Y}$ then by 7.8.1, $\phi_n(\mathbf{a}) \to \phi(\mathbf{a})$ for every $\mathbf{a} \in \mathbb{R}^k$, in particular for $\mathbf{a} = \mathbf{a}$.

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are k-dimensional random vectors with $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$, then the matrix $\mathbf{X}$ is the random $n \times k$ matrix (X_{ij}), where $X_{ij} = P(X_{ij} \leq x)$. The covariance matrix of $\mathbf{X}_i$ is the $k \times k$ matrix $\sum$, where the ij entry is $\mathrm{Cov}(X_{ri}, X_{rj})$. If the $\mathbf{X}_i$ are iid, we may speak of the common mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and covariance Σ. The following result gives some basic properties of the k-dimensional Gaussian distribution, which is discussed in Appendix F.

7.8.23 *A k-Dimensional Central Limit Theorem.* Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be iid k-dimensional random vectors with finite means and covariances Σ. If $\mathbf{S}_n = \sum_{i=1}^{n} \mathbf{X}_i$, then $\dfrac{\mathbf{S}_n - n\boldsymbol{\mu}}{\sqrt{n}}$ converges weakly to $\mathbf{Y}$, where $\mathbf{Y}$ has a Gaussian distribution with mean $\mathbf{0}$ and covariance Σ.

Proof. Let $\mathbf{Y}$ have a Gaussian distribution with mean $\mathbf{0}$ and covariance Σ. (Such a random vector exists since Σ is symmetric and nonnegative definite.) By the Cramér-Wold device, it is sufficient to show that for every $\mathbf{a} \in \mathbb{R}^k$,

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{a} \cdot \mathbf{X}_i - \mathbf{a} \cdot \boldsymbol{\mu}) \xrightarrow{d} \sum_{i=1}^{k} a_i Y_i.$$

But the random variables $Y_i = \sum_{j=1}^{k} a_j X_{ij}$ are iid with mean $\sum_{j=1}^{k} a_j \mu_j$ and variance $\sum_{r=1}^{k} \sum_{s=1}^{k} a_r a_s \sigma_{rs}$. [Note that $\sigma_{rs} = \mathrm{Cov}(X_{1r}, X_{1s})$.

Now,

$$T_n = \frac{\sum_{i=1}^{n} (X_i - \theta)(Z_i)}{\sqrt{n}}$$

and the one-dimensional central limit theorem says that T_n converges weakly to a normal distribution with mean 0 and variance $\sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j \sigma_{ij}$ say. But this is precisely the distribution of $\sum_{i=1}^{k} a_i Y_i$. ∎

7.9 Bibliography

An extensive treatment of the one-dimensional central limit theorem is given in Gnedenko and Kolmogorov (1954). Conditions are derived for convergence in distribution of the row sums of a triangular array to a given infinitely divisible distribution. In Section 6 a given stable random variable with distribution function G and $X_1, X_2, \ldots$ are iid random variables with distribution function F, conditions on F are given that are necessary and sufficient for there to exist a sequence of normalizing constants $\{a_n\}$ and $\{b_n\}$ such that $a_n(X_1 + \cdots + X_n) - b_n$ converges in distribution to G. In this case F is said to belong to the domain of attraction of G. Also, results on moment convergence are given; in other words, if $Y_n \xrightarrow{\mathcal{L}} Y$, then the conditions are given that the distribution function of Y_n approximates that of Y.

Prokhorov's weak compactness theorem actually holds in a complete separable metric space, see Billingsley (1968) for a proof. An n-dimensional extension of Lindeberg's theorem is given by Gikhman and Skorokhod (1969). Aspects of the central limit theorem for random variables with values in abstract spaces, for example, Hilbert spaces or locally compact groups, are discussed in Parthasarathy (1967).

For further applications in statistics see Ferguson (1967) and Sverling (1980).

CHAPTER

8

ERGODIC THEORY

8.1 Two examples

In Chapter 6 we proved the strong law of large numbers: if $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent identically distributed random variables with finite mean, then

$$\frac{X_1 + \cdots + X_n}{n} \to \mathbb{E}(X_1) \quad \text{a.e.}$$

In this chapter we will generalize this result and prove the more "primitive" ergodic theorem.

The starting point for ergodic theory is the notion of a transformation that preserves the structure of the measure space, as defined below.

8.1.1 Definition Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space and T a measurable transformation on $(\Omega, \mathcal{F}, \mu)$, that is, $T^{-1}(\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$.

The transformation T is said to be *measure preserving* with respect to μ if μ is preserved by T, that is, $\mu(A) = \mu(T^{-1}(A))$ for each $A \in \mathcal{F}$.

This implies that $\mu(T^{-k}A) = \mu(A)$ for all $A \in \mathcal{F}$ and all $k = 1, 2, \ldots$, where $T^{-k}A = \{x : T^k x \in A\}$ and T^k is the composition of T with itself k times.

The physical concept of a *flow* may be used to motivate the study of measure preserving transformations. A flow may be regarded as a process in which a system of particles in a fluid (each point of the container corresponding to a particle) moves about under the action of an externally applied force. The flow is assumed to be independent of time, so that, if beginning observations at $t = 0$, 1, 2, $\ldots$, the flow can be described by a single measurable function T. If x is a point of the container, Tx is the position the particle originally at x will occupy after one second has elapsed; thus $T^2x = T(Tx)$ is the position after two seconds, and so on. If A is a region (Borel) in the container,

that $T^{-1}\{i\}$ corresponds to the set of particles that will be in area i after one application of T. Then μ is Lebesgue measure on Ω in this case, and "the limit is incomprehensible, it is reasonable to expect that $\mu(T^{-1}\{i\}) = \mu(\{i\})$.

We consider some mathematical examples.

8.1.2 Examples

1 *Permutations.* Let Ω be a finite set $\{x_1, \ldots, x_n\}$, $n \geq 2$, with $\mathcal{F}$ consisting of all subsets of Ω. Let T be a cyclic permutation of Ω, say $Tx_i = x_{i+1}$ with indices taken modulo n. Since $T^{-1}\{x_i\} = \{x_{i-1}\}$, T preserves μ if $\mu(x_i) = \mu(x_{i-1})$ for all i. Thus if μ is a probability measure, then $\mu\{x_i\}$ must be $1/n$ for all i.

More generally, if T is any permutation of Ω, T can be expressed as a product of disjoint cycles (see Exercise 1.2.7). Then T preserves μ if within each cycle μ assigns equal weight to each point.

2 *Translations.* Let $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, and μ be Lebesgue measure. If $Tx = x - a$ for a constant a, then T preserves μ, because μ is translation-invariant.

3 *Rotations of the circle.* Let Ω be the unit circle in the plane $\mathbb{R}^2$; Ω can be identified with the interval $[0, 2\pi)$ under the correspondence $x^2 + y^2 = 1$ $\leftrightarrow$ θ, where θ is the length of the arc on the circle. Let Lebesgue measure on $[0, 2\pi)$ be denoted by μ, and let $\mathcal{A}$ be a Borel subset of $[0, 2\pi)$; $\mu(\mathcal{A}) = \int_{\mathcal{A}} (1/2\pi)\, d\theta$.

Let α be fixed in $[0, 2\pi)$, and let T be rotation by α. Thus T is defined on Ω by $T(e^{i\theta}) = e^{i(\theta+\alpha)}$, or equivalently on $[0, 2\pi)$ by $T(\theta) = \gamma = \theta + \alpha \pmod{2\pi}$. As in Example 2, T preserves μ by the translation invariance of Lebesgue measure.

4 *One-sided shifts.* Let $\Omega = A^{\mathbb{N}}$, the collection of all sequences $x = (x_0, x_1, \ldots)$ of real numbers. Let $\mathcal{F} = \mathcal{B}(\mathbb{R}^{\mathbb{N}})$, and let μ be any probability measure on $\mathcal{F}$. Define $T(x_0, x_1, \ldots) = (x_1, x_2, \ldots)$; T is called the *one-sided shift transformation*. Measures preserved by T are stationary in the sense of Definition 8.1.1 below.

5 *Two-sided shifts.* Let Ω be the set of all doubly infinite sequences $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ of real numbers, $\mathcal{F}$ the σ-field generated by the measurable cylinders

$$\{x : (x_n, \ldots, x_{n+k}) \in B_{n,k}\}$$

$n = \ldots, -2, -1, 0, 1, 2, \ldots$, $k = 0, 1, 2, \ldots$, $B_{n,k} \in \mathcal{B}(\mathbb{R}^{k+1})$. Let μ be any probability measure on $\mathcal{F}$, and let T be the *two-sided shift* defined by

$$T(\ldots, x_{-1}, x_0, x_1, \ldots) = (\ldots, x_0, x_1, x_2, \ldots).$$

In other words, if the kth coordinate axis is x_k, the kth coordinate of Ty is x_{k+1}. As in Example 4, m is invariant under T and stationary in the sense of Definition 5.1.2.

In Example 4, the coordinate variables are defined as follows. If $x = (x_0, x_1, x_2, \ldots)$, then $X_k(x) = x_k$, $k = 0, 1, \ldots$. (A similar definition is made in Example 5.) If T is the one-sided shift on Example 5, m, the one-sided shift, we have

$$X_k(Tx) = X_{k+1}(x).$$

5.1.3 *Definition.* Let P be a probability on $\mathbb{R}^\infty$. P is stationary if

$$P[x : (x_0, \ldots, x_{k-1}) \in B_k] = P[x : (x_1, \ldots, x_k) \in B_k]$$

for all $k = 1, 2, \ldots$, and all k-dimensional Borel sets B_k.

In the case of doubly infinite sequences, $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$, it is optional for $k = \ldots, -1, 0, 1, \ldots$.

We show that T preserves P if P is stationary.
Indeed, note that

$$P^{-1}[x : (x_0, \ldots, x_{k-1}) \in B_k] = P[x : (x_1, \ldots, x_k) \in B_k]$$

If T preserves P, then

$$P(B) = P(T^{-1}B) = \cdots = P(T^{-n}B) = \cdots$$

and it is easy to see that P is stationary.

Conversely, if P is stationary and $A = [x : (x_0, \ldots, x_{k-1}) \in B_k]$ is a cylinder set, then $T^{-1}A = [x : (x_1, \ldots, x_k) \in B_k]$. The class of sets $A \in \mathcal{F}$ such that $P(A) = P(T^{-1}A)$ is a monotone class containing the measurable cylinders, and hence coincides with $\mathcal{F}$. The result follows.

5.1.4 *Definition.* Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X_0, X_1, \ldots$, a sequence of random variables. The sequence $X_0, X_1, \ldots$ is a stationary sequence if for each $k = 0, 1, \ldots$ and each $n = 1, 2, \ldots$, $(X_0, \ldots, X_{n-1})$ and $(X_k, \ldots, X_{k+n-1})$ have the same distribution.

In the case of doubly infinite sequences of random variables $\ldots, X_{-1}, X_0, X_1, \ldots$, we shall assume by $k = \ldots, -1, 0, 1, \ldots$.

If $X_0, X_1, \ldots$ is a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, we can define a measure μ on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ a probability space as follows: if B_n is a Borel set in $\mathbb{R}^n$, we take

$$\mu[x : (x_0, \ldots, x_{n-1}) \in B_n] = P[\omega : (X_0(\omega), \ldots, X_{n-1}(\omega)) \in B_n].$$

This probabilities μ_n are consistent in the sense of the Kolmogorov extension theorem (2.7.3), and define a probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ characterized by

$$P\{(x_0, x_1, \ldots) : (x_k, x_{k+1}, \ldots) \in B\} = \mu_k(B), \quad B \in \mathcal{B}(\mathbb{R}^{k+1}).$$

It is easy to verify that the sequence $X = (X_0, X_1, \ldots)$ is stationary if and only if the probability P is a stationary probability on $\mathbb{R}^{\mathbb{N}}$.

Note that the transformations of Examples 1, 2, 3, and 5 are invertible measure preserving mappings, while T in Example 4 is measure preserving, while that in Example 4 is neither invertible nor measure preserving.

We now consider a physical example to illustrate the concept of an ergodic transformation. Suppose that rainfall data are collected at a very large number of observation points (x, y), $x = 0, 1, \ldots$, $y = 0, 1, \ldots$. Assume that the qualitative character of the observations at (x, y) is the same for each (x, y). The phenomenon can be represented by a stationary sequence of random variables $X_{11}, X_{12}, \ldots$, with X_{ij} the amount of rainfall at time i at site j. Assume that the X_{ij} are "nonpathological," in other words that the sequence is a sequence of independent performances of a random experiment, where a performance consists in observation of the entire sequence $(X_{ij}, X_{i+1,j}, \ldots)$.

Suppose that the problem is to measure the average rainfall. Scientist A might take the following approach. He or she might take measurements at each observation point at a given time, say $t = 0$, and average the results. Scientist B might reason as follows. Since all observation points have the same statistical behavior, we can simply go to one observation point, take a large number of observations, say at $t = 0, 1, \ldots, n - 1$, and average the results. Scientist A is using what might be called a vertical measuring scheme, and Scientist B a horizontal scheme. In the language of the ergodic theorem, A's observations correspond to the first column, B's to the first row.

	Measurements			
Observation point	$\cdots$	1	0	$\cdots$
0	$X_{-1,0}$	X_{01}	X_{00}	$\cdots$
1	X_{10}	X_{11}	X_{10}	$\cdots$
0	X_{00}	X_{01}	X_{02}	$\cdots$
$\vdots$				

Now A and B will not necessarily obtain the same results (but "reasonably" the same). For example, suppose that "nature" flips an unbiased coin at each observation point. If the coin comes up heads, the rain is one inch at each observation time; if tails, there is no rain at any time. Roughly half of A's measurements will measure one inch of rainfall, and half will

...

$$\frac{1}{n}\sum_{i=1}^{n-1} Y_i(\omega) \qquad \text{for a permutation } \sigma$$

...

But A is observable on countable intervals of a particular kind, namely, $I_i(\omega) \sum_{j=1}^{n} Y_j$ because the $Y_i = X_i$ are independent random variables, all having the same distribution as $Y_1 = X_1$. Thus A's mean would approximate $E(X) = \int_\Omega X_1 dP$. For A and B to get the same answer we must have

$$\frac{1}{n}\sum_{i=1}^{n} X_i(T^i\omega) \to \int_\Omega X_1 dP.$$

More generally, we might ask when we can be sure that, for each integrable function f on $(\Omega, \mathcal{F}, P)$, we have

$$\frac{1}{n}\sum_{i=0}^{n-1} f(T^i\omega) \to \int_\Omega f dP \qquad (1)$$

at least for almost every ω. In particular, if F is an indicator for the property to be tested, then the convergence of the relative frequency of visits to A to the limit $\int_\Omega F dP$ is the probability of A.

Now suppose that A is an "invariant set" for the mapping T, and $T^{-1}A = A$. The only way a set can remain invariant under repeated flips of an unbiased coin is for the outcome to be all heads or all tails. Thus $A = \{\omega : X_1(\omega) = 1\}$ or $A = \{\omega : X_1(\omega) = 0\}$. Then we have almost everywhere

$$\frac{1}{n}\sum_{i=0}^{n-1} I_A(T^i\omega) = I_A(\omega) \qquad \text{for all } n.$$

Thus the relative frequency of visits to A cannot converge to $P(A) = \frac{1}{2}$ except when $P(A) = 0$ or 1. Conversely, the pointwise ergodic theorem, to be proved in §5.3, implies that if every invariant set has probability 0 or 1, the convergence result in (1) holds for the next section we shall prepare for the proof of this basic result.

One comment on terminology. In this section we have used random variable with real values. We employ this terminology

8.2 Recurrence and Mixing

The following definitions are motivated by the analysis at the end of 8.1.

8.2.1 *Definition.* Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$. A set $A \in \mathcal{F}$ is said to be *invariant* if $T^{-1}A = A$, that is, $\omega \in A$ iff $T\omega \in A$. Two sets A and B are equivalent if $T^{-1}A$ and B differ by a set of measure 0, in other words, $\mu(A \triangle T^{-1}A) = 0$.

It is easily checked that the invariant sets form a σ-field, as do the almost invariant sets.

If $(S, \mathcal{S})$ is a measurable space and $g : \Omega \to S$ is measurable, then g is said to be *invariant* if $g \circ T = g$ a.e. and *almost invariant* if $g \circ T = g$ a.e. Note that a set is invariant (respectively almost invariant) iff its indicator is invariant (almost invariant).

The measure preserving transformation T is said to be *ergodic* if for every invariant set A, either $\mu(A) = 0$ or $\mu(A^c) = 0$. In the case of a probability space, ergodicity means that each invariant set has probability 0 or 1.

Invariance may be replaced by almost invariance in the definition of ergodicity, as the following result shows.

8.2.2 *Lemma.* Let T be a measure preserving transformation.

(a) If A is an almost invariant set, there is a strictly invariant set B such that $\mu(A \triangle B) = 0$.

(b) A measure-preserving transformation T is ergodic if T for each almost invariant set A, either $\mu(A) = 0$ or $\mu(A^c) = 0$.

Proof. Take $B = \limsup_n T^{-n}A$. Then $T^{-1}B = \limsup_n T^{-(n+1)}A = B$, hence B is invariant.

Now $A \triangle B \subset \bigcup_{n=0}^{\infty} T^{-n}A \triangle T^{-n-1}A$, and for each n, $\mu(T^{-n}A \triangle T^{-n-1}A) = 0$ since T^{-n} is measure preserving. Hence $\mu(A \triangle B) = 0$. Finally, $T^{-n}A \supset B$ for all n, so $T^{-n}A \supset T^{-n}B \supset T^{-n-1}B$. Thus $T^{-n}B$ is the smallest integer such that $T^{-n}B \supset A$, that is, $T^{-n}B = T^{-n}A$.

Since $\mu(T^{-n}A \triangle B) = \mu(A \triangle T^{n}B) = 0$, it follows that $\mu(A \triangle B) = 0$, proving (a).

Suppose (b) holds and let A be a set with $T^{-1}A = A$. If B is as above and $\mu(A \triangle B) = 0$, then $\mu(B) = \mu(A) = \mu(A \cap B) + \mu(A \setminus B) = \mu(A \cap B) = \mu(B) \cdot \mu(B)$. Hence $\mu(B) = 0$ or $\mu(B^c) = 0$. The converse is clear since every invariant set is almost invariant. □

We give another way of expressing ergodicity.

4.2.3 Lemma. *Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$. The following conditions are equivalent:*
(a) T is ergodic;
(b) Every almost invariant function is a.s. constant;
(c) Every invariant function is a.s. constant.

Proof. *(a) $\Rightarrow$ (b).* Let u be an almost invariant function. Then for each real λ, $A_\lambda = \{\omega : u(\omega) \le \lambda\}$ is an almost invariant set since $\mu(T^{-1}A_\lambda \Delta A_\lambda) = 0$ or $\mu(A_\lambda \Delta T^{-1}A_\lambda) = 0$. Thus $\mu(A_\lambda) = 0$ or 1. The only possibility is $\mu(A_\lambda) = 0$ for $\lambda < \alpha$ and $\mu(A_\lambda) = 1$ for $\lambda > \alpha$, since $A_\lambda \uparrow \Omega$ as $\lambda \uparrow \infty$ and $A_\lambda \downarrow \emptyset$ as $\lambda \downarrow -\infty$. Then

$$\mu\{\omega : u(\omega) = \alpha\} = \mu\left[\bigcap_{n=1}^{\infty} \{\omega : \alpha - \tfrac{1}{n} < u(\omega) \le \alpha\}\right] = 0,$$

and similarly $\mu\{\omega : u(\omega) = \alpha^+\} = 0$. Thus $u = \alpha$ a.s.

(b) $\Rightarrow$ (c). Every invariant function is almost invariant.

(c) $\Rightarrow$ (a). If A is an invariant set, χ_A is an invariant function hence is a.s. constant. If $\chi_A = 0$ a.e. then $\mu(A) = \int_\Omega \chi_A(\omega)\,d\mu = 0$, and if $\chi_A = 1$ a.e. then $\mu(A^c) = \int_\Omega (1 - \chi_A)\,d\mu = 0$. $\square$

Invariant and almost invariant random variables are defined in the same way as invariant and almost invariant Borel measurable functions. Lemma 4.2.3 holds in this case also, with essentially the same proof.

The following characterization of almost invariant sets is often useful.

4.2.4 Lemma. *Let T be a measure-preserving transformation. Assume that μ is finite. A set $A \in \mathcal{F}$ is almost invariant if and only if there exists $B \in \mathcal{F}$ such that $T^{-1}B = B$ and $\mu(A \Delta B) = 0$; that is, T is essentially implied by A and A is essentially implied by T.*

Proof. We have only

$$\begin{aligned}\mu(A \Delta T^{-1}A) &= \mu(A \cap T^{-1}A^c) + \mu(A^c \cap T^{-1}A) \\ &= \mu(T^{-1}A^c) - \mu(A^c) \\ &= \mu(A^c) - \mu(A^c) = 0.\end{aligned}$$

Now apply the B of the existence of a nonsingular transformation, indicates that T does not completely set up the space. This concept of "mixing" must be developed as follows.

5.2.5 Definition. If T is a measure preserving transformation on the probability space $(\Omega, \mathcal{F}, P)$, T is said to be *mixing* if for all $A, B \in \mathcal{F}$,

$$\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B).$$

The normalization to a probability measure is essential here. If, for example, T has the mixing property with respect to the measure m on $\mathcal{F}$ and $A = B = \Omega$ in (5.2.5) then $m(\Omega) = [m(\Omega)]^2$. Thus if $m(\Omega)$ is finite it must be 1; if $m(\Omega) = \infty$, and A is chosen so that $0 < m(A) < \infty$, then $m(A \cap T^{-n}A) = m(A) \neq m(A)m(A)$. Thus, provided $m(A)m(A) = m(A)$ is not a contradiction.

The following property has the following intuitive interpretation (this example is due to Halmos (1956)). Recall the transformation T as defining a flow, as in the discussion after 5.1.1. Suppose, for example, that in 1776 the contents of a bottle of perfume is poured into the corner of a room, the virgin particles of the perfume $= A$. The eventually applied force is due to a sizable stock. The condition of the container is observed at times $t = 0, 1, 2, \ldots$. If B is any Borel subset of the container, let $P(B)$ be the volume of B divided by the volume of the container so that $P(A) = 0$ and $P(\Omega) = 1$. It is reasonable to assume that if the mixing process is continued long enough, the percentage of perfection in B should be approximately the same as the percentage of the entire container, namely, $100P(B)$. To translate this into mathematical terms note that if ω is a point of the container, and a particle is initially at ω, then $T^n\omega$ is the position of the particle a second later. Thus the set of momentum particles that are in B at time $t = n$ is $\{\omega : T^n\omega \in B\} = A \cap T^{-n}B$. The percentage of perfection in B at time $t = n$ is $P(A \cap T^{-n}B)/P(B)$, and the mixing property is expressed by saying that $P(A \cap T^{-n}B)/P(B) \to P(A) = 0$.

Mixing is a stronger property than ergodicity as we now prove:

5.2.6 Theorem. Let T be a mixing transformation on $(\Omega, \mathcal{F}, P)$. Then T is ergodic.

Proof. Let B be an invariant set. If $A = B$ then since $B = T^{-n}B$ we have $P(A \cap T^{-n}B) = P(B \cap T^{-n}B) = P(B)$ for all n. Taking the limit and invoking the mixing property we obtain $P(B) = P(B)P(B)$. Therefore $P(B) = 0$ or 1. Since A is an arbitrary set in $\mathcal{F}$, we must take $A = B$, and $P(B) = [P(B)]^2$. Hence $P(B) = 0$ or 1. $\square$

It is useful to observe that it is not necessary to verify the mixing condition for all the sets $A, B \in \mathcal{F}$, but only for A, B in a field $\mathcal{G}_0$ whose minimal σ-field is $\mathcal{F}$.

5.2.7 Theorem. Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, P)$. Let $\mathcal{G}_0$ be a field of subsets of Ω such that the minimal σ-field containing $\mathcal{G}_0$ is $\mathcal{F}$.

By 1.6.12,

$$\pi_k = \int_{[0,1)} e^{-2\pi i k x}\, d\mu(x) = e^{-2\pi i k/n} \int_{[0,1)} e^{-2\pi i k x}\, d\mu(x) = e^{-2\pi i k/n}\pi_k.$$

By hypothesis of the theorem $e^{-2\pi i k/n} \neq 1$ for $0 < k < n$. It follows that $\pi_k = 0$ for $k \neq 0$. Thus $f_\mu = f^\Sigma$, so that the Fourier series $\sum_{k=-\infty}^{\infty} \pi_k e^{2\pi i k x}$ converges in L^2 to $f_\mu(x)$ (see 5.2, Theorem 1). It follows that $f_\mu = a_1 = a_2 = \cdots$, and therefore $P(A) = 0$ or 1. Thus T is ergodic.

Conversely, if $e^{2\pi i \alpha} = e^{2\pi i p/q}$ is a root of unity, say, $e^{2\pi i \alpha q} = 1$, then α is an irrational multiple of $2\pi/q$. Let A be the union of the sets $[0, \frac{1}{2q}) \cup [\frac{2}{2q}, \frac{3}{2q}) \cup \cdots \cup [\frac{2q-2}{2q}, \frac{2q-1}{2q})$. Then $A \cap T^{-1}A = [0, \frac{1}{2q}) \cup [\frac{2}{2q}, \frac{3}{2q}) \cup \cdots \cup [\frac{2q-2}{2q}, \frac{2q-1}{2q})$. Hence A is invariant but $P(A) = \frac{1}{2}$, so that T is not ergodic.

Now T is easily seen to establish this we may assume that $\alpha/2\pi$ is irrational. Let $\alpha = \beta - [\beta]$. If $\alpha \leq \frac{1}{2}$, an inequality in the upper semicircle. Given $\varepsilon > 0$, $e^{i\beta}$ is within distance ε of $e^{2\pi i \gamma}$ for infinitely many n. [Choose a convergent $\frac{p}{q}$ to γ from $(\gamma, \gamma + \varepsilon)$, and notice $q\gamma - pq$ and $pq - q\gamma$ such that ε can be chosen larger than any preassigned positive integer. Then it is possible to form a chain of consecutive open intervals around the circle, each of length ε, with successive points less than ε apart, so that A and TA overlap except on a set of measure less than ε. Thus

$$P(A \cap T^{-1}A) = \int_{A \cap T^{-1}A} f(x)\, dx = P(T^{-1}A) = P(A),$$

since T is measure preserving and therefore

$$\Rightarrow P(A) = \frac{1}{2} = \varepsilon > \frac{1}{2} \quad \text{for} \quad \varepsilon < \frac{1}{2}.$$

Thus $P(A)[1 - P(A)] = P(A) \cdot P(A^c) = 0$, so that A is trivial.

One-sided and two-sided stationary processes. Let T be the one-sided (or two-sided) shift transformation on the space of all infinite (or doubly infinite) sequences of real numbers. We consider only the case in which the coordinate random variables X_k are independent, that is, the measure P has the property that

$$P\{\omega : X_k(\omega) \in A_k,\ k = 1, 2, \ldots, n\} = \prod_{k=1}^{n} P\{\omega : X_k(\omega) \in A_k\}$$

for all Borel sets $A_1, \ldots, A_n$ and all $n = 1, 2, \ldots$. In this case, T is ergodic because ergodicity (i.e.,

$$P\{\omega : (X_1(\omega), \ldots, X_n(\omega)) \in B_n,$$

and

$$\mathcal{B} = \sigma\{B_0 \times B_1 \times \cdots \times B_{n-1} \times [0,1] \times [0,1] \times \cdots : B_i \in \mathcal{B}_1\}$$

be measurable cylinders. For sufficiently large n we have $A = C \times [0,1] \times [0,1] \times \cdots$ for some $C \in \mathcal{B}_n$. Therefore A and $T^{-1}(B)$ are cylinders. Thus by independence,

$$P(A \cap T^{-1}(B)) = P\{(x_0, \ldots, x_{n-1}) \in C, (x_1, \ldots, x_n) \in B_0 \times \cdots \times B_{n-1}\}$$
$$= P(A)P(B_0) \cdots P(B_{n-1}) = P(A)P(B).$$

Thus the monotone class theorem implies $P(A \cap T^{-1}(B)) = P(A)P(B)$ for all measurable cylinders, and hence by 8.2.1, the relation holds for all $A, B \in \mathcal{A}$, so that T is mixing.

Problems

1. If $\mathcal{A}$ is an almost circular set, show that one cannot assign every a_n a value in $\{0, 1\}$ so that $T^n(a_n) \in A$ for all $n = 1, 2, \ldots$.
2. If μ is a rotating measure on the circle, show that $\nu = \frac{1}{2}(\mu + \mu^{-1})$ is symmetric, but $\nu^{*n} = \nu^{*(n-1)} * \nu$ is not.
3. Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$. Give examples to show that the following results are possible.
   (a) $A \in \mathcal{F}$ does not imply $T(\omega) \in \mathcal{F}$.
   (b) If $A \in \mathcal{F}$ and $T(\omega) \in \mathcal{F}$, $P(A)$ and $P(T(\omega))$ need not equal $P(\Omega)$.
   (c) If T is one-to-one onto, it need not be invertible (that is, T^{-1} need not be measurable).
4. (Jacobs, 1962) Let T be a measurable (but not necessarily measure-preserving) transformation on $(\Omega, \mathcal{F}, \mu)$. We say that T is *recurrent* iff for every $A \in \mathcal{F}$ and almost every $\omega \in A$, $T^n(\omega) \in A$ for infinitely many n. T is *infinitely recurrent* iff for every $A \in \mathcal{F}$ and almost every $\omega \in A$, $T^n(\omega) \in A$ for infinitely many n. (In these definitions, the exceptional set of measure 0 is allowed to depend on A.) A set $B \in \mathcal{F}$ is *wandering* if $B, T^{-1}B, T^{-2}B, \ldots$ are disjoint. T is said to be *conservative* if all wandering sets have measure 0. Finally, T is *incompressible* if $A \subset T^{-1}A$ implies $\mu(T^{-1}A - A) = \mu(T^{-n}A - T^{-(n-1)}A) = 0$. (Show that implies only $T^{-1}A \subseteq A$ implies $\mu(A - T^{-1}A) = 0$.)
   (a) Show that the following are equivalent.
      i. T is incompressible
      ii. T is conservative
      iii. T is recurrent
      iv. T is infinitely recurrent

   (Two possible schemes: (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv), (iv) $\Rightarrow$ (i); (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (iv) $\Rightarrow$ (i).)

8: Show that $T(x) = x + \alpha \bmod 1$ on $\mathbb{R}$ with Borel sets and Lebesgue measure violates all four conditions of (a).

9: (Poincaré) If T is measure-preserving and μ is finite, show that T is (infinitely) many-to-one.

8.3 The Pointwise Ergodic Theorem

We will prove the pointwise ergodic theorem, which says that if T is a measure-preserving transformation on $(\Omega, \mathcal{F}, \mu)$ and $f \in L^1(\Omega, \mathcal{F}, \mu)$, then $\frac{1}{n}(f(\omega) + f(T\omega) + \cdots + f(T^{n-1}\omega))$ converges to an integrable function $\bar{f}(\omega)$, for almost all ω.

It is possible to prove this result in a somewhat more general setting than the one we have at the end of the chapter. The generalization is based on the fact that the operation of taking a measure-preserving transformation is a positive contraction operator, as follows.

8.3.1 *Theorem.* If f is an extended real-valued Borel measurable function on $(\Omega, \mathcal{F}, \mu)$, and T is a measure-preserving transformation, and we define the function $f \circ T$ by $(f \circ T)(\omega) = f(T\omega)$, then for every $p \geq 1$, $\|f \circ T\|_p = \|f\|_p$. This is the reason we consider T as a linear transformation on the function spaces $L^p(\Omega, \mathcal{F}, \mu)$. For $1 \leq p < \infty$, then $\hat{T}$ is an isometry (preserves norms), linear, and positive (maps positive functions to positive functions). In particular, T is a contraction (that is, $\|\hat{T}\| \leq 1$) (in this case, $\|\hat{T}\| = 1$). Furthermore, $\hat{T}$ is positive, that is, if $f \geq 0$ a.e., then $\hat{T}f \geq 0$ a.e.

Proof. By (4.6.12), $\int_\Omega |f \circ T|^p \, d\mu = \int_\Omega |f|^p \, d\mu$ (take $g = 1_\Omega$ in the change of variables formula). Hence $\|\hat{T}f\|_p = \|f\|_p$, so $\hat{T}$ is a contraction. If $f = 0$ a.e., then $f \circ T = 0$ a.e., so $\hat{T}f = 0$ a.e. Thus $\hat{T}$ is positive. Finally, we have

$$\begin{aligned} \|\hat{T}^n f\|_\infty &= \inf\{a : \hat{T}^n f(\omega) \leq a \text{ a.e.}\} = 0\} \\ &= \inf\{a : f(T^n \omega) \leq a \text{ a.e.}\} \\ &= \inf\{a : \mu T^{-n}\{\omega : f(\omega) > a\} = 0\} \\ &= \inf\{a : \mu\{\omega : f(\omega) > a\} = 0\} \quad \text{since } T \text{ preserves } \mu \\ &= \|f\|_\infty \end{aligned}$$

If $p = \infty$, $f \geq 0$ a.e., and $g = f^{1/n}$, then $g \geq 0$ a.e. and $\hat{T}f = (\hat{T}g)^n \geq 0$ a.e., proving positivity.

The sequence of functions $f(\omega), f(T\omega), f(T^2\omega), \ldots, f(T^{n-1}\omega)$ can now be expressed in terms of $\hat{T}$ as

$$f, \hat{T}f, \hat{T}^2 f, \ldots, \hat{T}^{n-1} f$$

where $\hat{T}^0 f = f$ and $\hat{T}^k$ is the composition of $\hat{T}$ with itself k times, $k \geq 1$.

In the three exercises that follow, T is a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$, and $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Inspection of the proofs will show, however, that the results hold if T is replaced by an arbitrary positive contraction on L^1 not necessarily arising from a measure-preserving transformation.

8.3.2 *Lemma.* If $n \in \mathbb{N}$, $f \in L^1(\Omega, \mathcal{F}, \mu)$, and $f_n = \max(0, f, f + Tf, \ldots, f + Tf + \cdots + T^{n-1}f)$, then

$$\int_{\{f_n > 0\}} f \, d\mu \geq 0.$$

Thus

$$\int_{\{F_n > 0\}} f \, d\mu \geq \int_{\{F_n > 0\}} Tf \, d\mu, \qquad n = 0, 1, \ldots.$$

Proof. If $\omega \in \{f_n > 0\}$, then $\sum_{k=0}^{n-1} T^k f = f + T\left(\sum_{k=0}^{n-1} T^k f\right)$. Since $\sum_{k=0}^{n-1} T^k f \leq f_n$ and T is positive, we have

$$T\left(\sum_{k=0}^{n-1} T^k f\right) \leq Tf \, 1_{\{F_n > 0\}} \leq Tf^+.$$

Thus

$$\sum_{k=1}^{n} T^k f \leq f + Tf^+, \qquad \omega \in \{F_n > 0\}.$$

Since $f \leq f + Tf^+$, we have $f_n \leq f + Tf^+$ on $\{F_n > 0\}$.

8.3.3 *Lemma.* Let f_n be defined as in 8.3.2 and assume $f \in L^1(\Omega, \mathcal{F}, \mu)$. If $A = \{f_n > 0\}$, then $\int_A f \, d\mu \geq 0$.

Proof. By 8.3.2,

$$\begin{aligned}
\int_{A_n} f \, d\mu &\geq \int_{A_n} f 1_{\{F_n > 0\}} \, d\mu - \int_{A_n} (Tf^+) 1_{\{F_n > 0\}} \, d\mu \\
&\geq \int_{A_n} f_n \, d\mu - \int_{A_n} (Tf^+) 1_{A_n} \, d\mu \\
&= \int_{A_n} f_n^+ \, d\mu - \int_{A_n} (Tf^+) 1_{A_n} \, d\mu \\
&\geq \int_{A_n} f_n^+ \, d\mu - \int_{A_n} (Tf_n^+) \, d\mu \\
&= \|f_n^+\|_1 - \|Tf_n^+\|_1 \geq 0
\end{aligned}$$

since T is a contraction. $\square$

8.3.4 *Maximal Ergodic Theorem.* If $f \in L^1(\Omega, \mathcal{F}, \mu)$ and

$$A = \left\{\omega : \sup_{n \geq 1} \sum_{k=0}^{n-1} f(T^k \omega) > 0\right\}$$

$$= \left\{\omega : \sup_{n \geq 1} f_n^+(\omega) > 0\right\} \quad \text{where} \quad f_n^+ = \max\left\{0, \sum_{k=0}^{n-1} f \circ T^k\right\}$$

then $\int_A f \, d\mu \geq 0$.

Proof. The sets A_n of (8.3) increase to A.

I will first comment on Garsia's proof of the technical difficulties in the proof of the pointwise ergodic theorem. Let $f \in L^1(\Omega, \mathcal{F}, P)$ and let $\bar{f}(\omega) = \limsup_{n \to \infty} \frac{1}{n} \sum_{m=0}^{n-1} f(T^m \omega)$, $n = 1, 2, \ldots$, where T is measure preserving. If $\alpha < \beta$ define

$$G_{\alpha\beta} = G_{\alpha\beta}(f) = \left\{\omega : \liminf_{n \to \infty} f^{(n)}(\omega) < \alpha < \beta < \limsup_{n \to \infty} f^{(n)}(\omega)\right\},$$

$$N_\alpha = \left\{\omega : \sup_{n \geq 1} f^{(n)}(\omega) > \alpha\right\}.$$

Note that $G_{\alpha\beta}$ is a subset of N_α.

Instead of (8.3) we can establish the convergence of the sequence $f^{(n)}$ if we show that $P(G_{\alpha\beta}) = 0$ for all $\alpha < \beta$. We may assume without loss of generality that $\alpha > 0$. Note that

$$E|f| = \int |\omega : \liminf f^{(n)}(\omega) < \alpha < \beta < \limsup f^{(n)}(\omega)| \, dP$$

$$= E(|f| \mid G_{\alpha\beta}) P(G_{\alpha\beta}).$$

If $\beta \geq 0$, then $G_{\alpha\beta} \subset N_\alpha$ and the argument below will show that $E(|f| \mid G_{\alpha\beta}) < 0$ for measure 0 until α reaches 0. Hence $P(G_{\alpha\beta}) = 0$.

8.3.5 *Lemma.* The set $G_{\alpha\beta}$ has the following properties:

(a) The set $G_{\alpha\beta}$ is almost invariant.

(b) $\mu(G_{\alpha\beta}) < \infty$.

(c) If $\mu(G_{\alpha\beta}) > 0$

PROOF (1): We may write

$$\begin{aligned} P^{(n)}(T\omega) &= \frac{1}{n}\sum_{i=0}^{n-1} f(T^{i+1}\omega) \\ &= \frac{n+1}{n}\left[\frac{1}{n+1}\sum_{i=0}^{n} f(T^i\omega)\right] - \frac{f(\omega)}{n} \\ &= \frac{n+1}{n} P^{(n+1)}(\omega) - \frac{f(\omega)}{n}. \end{aligned}$$

Since $f \in L^1$, (1) is true for almost all ω, hence

$$\lim_{n\to\infty} P^{(n)}(\omega) = \lim_{n\to\infty} P^{(n)}(T\omega)$$

and similarly for $\lim \inf$, except possibly on a set of measure 0. Thus, outside a set of measure 0, $\lim \inf P^{(n)}(\omega) = \lim \sup P^{(n)}(\omega)$ iff $P^{(n)}(\omega) = P^{(n)}(T\omega)$, and the result follows.

(2): Let C be any set in $\mathcal{F}$ such that $C \subset \Gamma_\alpha$ and $\mu(C) < \infty$, and define $F_n = \{\omega : \sup_{k \geq n} |f(T^k\omega) - M_k(\omega)| > 0\}$. Note that if $\omega \in \Gamma_\alpha$, then $\omega \in N_\alpha$ for some n

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k\omega) > \alpha \quad \text{and hence} \quad \frac{1}{n}\sum_{k=0}^{n-1} M_k(\omega) > \alpha.$$

Thus $\omega \in F_n$ in particular, C is a subset of F_n. By the maximal ergodic theorem 8.3.4,

$$\int_{F_n} (f - M_n)\, d\mu \geq 0.$$

Thus

$$\int_{F_n} f\, d\mu \geq \int_{F_n} f\, d\mu + \alpha \int_{F_n} f_1\, d\mu - \int_{F_n} M_n\, d\mu \geq \alpha\mu(F_n) - \alpha\mu(C)$$

so that

$$\mu(F_n) \leq \frac{1}{\alpha}\int_{F_n} f\, d\mu < \infty.$$

But note that F_n is a subset of

$$\bigcup_{k=n}^{\infty} \{\omega : |f(T^k\omega)| > \alpha\}$$

By Theorem 1.6.12,

$$\int_{\mathbb{R}^n} |f(T^n x)|\,dx = \int_{\mathbb{R}^n} |f(x)|\,dx < \infty$$

and it follows that $|f(T^n x)| < \infty$ for a suitable subset of full measure, and therefore so is $f \circ T$ (see 8.2 Problem 2). Thus

$$\mu(C_{f,\alpha}) = \sup\{\mu(C) : C \in \mathcal{F}^+,\ C \subset C_{f,\alpha},\ \mu(C) < \infty\}$$

$$\leq \frac{1}{\alpha} \int_{\mathbb{R}^n} |f|\,d\mu < \infty.$$

(ii) Since $C_{f,\alpha}$ is almost invariant, T is a well-defined measure-preserving transformation on $(C_{f,\alpha}, \mathcal{F}_{C_{f,\alpha}}, \mu_{C_{f,\alpha}})$ where

$$\mathcal{F}_{C_{f,\alpha}} = \{B \cap C_{f,\alpha} : B \in \mathcal{F}\},$$

and $\mu_{C_{f,\alpha}} = \mu$ restricted to $\mathcal{F}_{C_{f,\alpha}}$. (Strictly speaking, if $D = C_{f,\alpha} \setminus T^{-1}C_{f,\alpha}$ then T is well-defined on $C_{f,\alpha} \setminus D$. Since $\mu(D) = 0$ this causes no difficulty. For example, we may regard T as the identity on D, and then it will be well-defined and measure preserving on $C_{f,\alpha}$.)

The ergodic theorem may now be applied to T on $C_{f,\alpha}$. In particular, the sequence $\frac{1}{n} S_n f = \frac{1}{n}(f + f \circ T + \cdots + f \circ T^{n-1})$ converges a.e. on $C_{f,\alpha}$ to a limit function f^* with

$$\int_{C_{f,\alpha}} |f^*|\,d\mu \leq \int_{C_{f,\alpha}} |f|\,d\mu < \infty.$$

Since $C_{f,\alpha}$ has finite measure, we may set $C = C_{f,\alpha}$ and since $C_{f,\alpha} \in \mathcal{F}_C$ we obtain

$$\int_C |f^*|\,d\mu \leq C.$$

Now let

$$A_{f,\alpha} = \left\{x \in C_{f,\alpha} : \limsup_{n} \frac{1}{n} S_n f(x) > \alpha\right\}.$$

If $x \in C_{f,\alpha}$ then $f^*(x) > \alpha$ for at least one n, hence $x \in T^{-n}C_{f,\alpha}$ and $C_{f,\alpha} = A_{f,\alpha}$. Since $C_{f,\alpha}$ has finite measure, constant functions are integrable on $C_{f,\alpha}$ and we may therefore apply the mean ergodic theorem to obtain

$$\int_C |f - f^*|\,d\mu = 0$$

Thus we have, for $n > 0$,

$$\lambda_n(C_n) \leq \int_{C_n} f \, d\mu \leq a\mu(C_n).$$

A contradiction follows as $\mu(C_n) = C_n$.

We may now prove the main result

8.16 *Birkhoff Ergodic Theorem.* Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, \mu)$, and let $f \in L^1(\Omega, \mathcal{F}, \mu)$. Then there is a function $f^* \in L^1$ such that we have $n^{-1} \sum_{k=0}^{n-1} f(T^k \omega) \to f^*(\omega)$ almost everywhere.

PROOF. Let $D = \{\omega : f^*(\omega)$ does not converge to a finite or infinite limit$\}$. Then $D = \bigcup_{r,s} D_{r,s}$, $r < s$, r, s rational. By 8.3.3, $\mu(D) = 0$, and hence $f^*(\omega)$ converges for almost all ω, we call the limit $f(\omega)$. Define $f = 0$ on the exceptional set. By Fatou's lemma,

$$\int_\Omega |f| \, d\mu = \int_{\Omega-D} \liminf_{n\to\infty} \left| \frac{S_n}{n} \right| d\mu \leq \liminf_{n\to\infty} \int_\Omega \left| \frac{S_n}{n} \right| d\mu \leq \int_\Omega |f| \, d\mu.$$

But

$$\begin{aligned} \int_{\Omega_a} |f^{(n)}| \, d\mu &= \frac{1}{n} \sum_{i=0}^{n-1} \int_{\Omega_a} |f \circ T^i| \, d\mu \\ &= \frac{n-a}{n} \sum_{i=0}^{n-a-1} \int_{\Omega_a} |f \circ T^i| \, d\mu \qquad \text{by 8.3.2} \\ &\to \int_{\Omega_a} |f| \, d\mu \end{aligned}$$

and the theorem is proved. □

We now look more closely at the convergence of the sequence $\{f^{(n)}\}$.

8.17 *Theorem.* If $\sup_n \{|f^{(n)}|\} \in L^p$ and $f^{(n)} \to f$ a.e., where $p \geq 1$, then $f \in L^p$ and $f^{(n)} \xrightarrow{L^p} f$.

PROOF. Since the finite-valued simple functions are dense in L^1, for each $\varepsilon > 0$ there is a bounded measurable function g such that $\|f - g\|_1 < \varepsilon$. If $\bar{f}(\omega) = f(T^n\omega)$, $\bar{g}(\omega) = g(T^n\omega)$, and $\|g\| \le K$, then

$$\left\| \frac{1}{n}\sum_{i=0}^{n-1} f(T^i\omega) - \bar{f}(\omega) \right\|_1 \le \left\| \frac{1}{n}\sum_{i=0}^{n-1} g(T^i\omega) - \bar{g}(\omega) \right\|_1 + \left\| \frac{1}{n}\sum_{i=0}^{n-1} f(T^i\omega) - \frac{1}{n}\sum_{i=0}^{n-1} g(T^i\omega) \right\|_1 + \|\bar{g} - \bar{f}\|_1 \quad (1)$$

Since $\left|\frac{1}{n}\sum_{i=0}^{n-1} g(T^i\omega)\right| \le K$ (since $|g| \le K$) and $\bar{g} \to \bar{f}$ a.e., the second term on the right in (1) approaches zero as n approaches ∞, by the dominated convergence theorem. (The hypothesis that $\mu(\Omega) < \infty$ implies that the constant function K is integrable.) By (1), $\|\bar{f}_n - \bar{g}_n\|_1 = \|f - g\|_1 < \varepsilon$, hence the first term is less than ε. Now

$$\int_\Omega \left| f - \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \right| d\mu = \int_\Omega \left| \frac{1}{n}\sum_{i=0}^{n-1} (f - g) \circ T^i \right| d\mu$$

$$+ \int_\Omega \left| \frac{1}{n}\sum_{i=0}^{n-1} g \circ T^i - \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i \right| d\mu$$

$$+ \left\| \frac{1}{n}\sum_{i=0}^{n-1} g \circ T^i - g \right\|_1 + \|g - f\|_1$$

Thus $\|f^{(n)} - f\|_1 < 3\varepsilon$ for large enough n, and the theorem follows. □

If $\mu(\Omega) = \infty$ in 8.3.1, the hypothesis that $\inf\{\|f - g\|_1 : g \text{ is invariant}\} = 0$ (see Problem 2) may be restrictive for $p = \infty$, even if $\mu(\Omega) = \infty$ (see Problem 3).

We can now identify the limit function $\bar{f}$. Theorem 8.3.1 implies that although the pointwise ergodic theorem can be presented without reference to probability, some insight is lost in doing so.

8.3.6 *Corollary.* If $T : \Omega \to \Omega$ is ergodic, then $\bar{f}$ is almost invariant. Thus if μ is a probability measure, $\bar{f} : (\Omega, \mathcal{H}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is constant a.e.

PROOF. If $A \in \mathcal{H}(T)$, then $\bar{f}(\omega) = \bar{f}(T\omega)$ a.e. on A and $\bar{f}(\omega) = \bar{f}(T\omega)$ a.e. on $\Omega - A$, so $\bar{f}(\omega) = \bar{f}(T\omega)$ for $\omega \in T^{-1}A$ where $\mu(T^{-1}A - A) = 0$. But by the proof of 8.3.5(a),

$$\bar{f}(T\omega) = \frac{n+1}{n}\bar{f}(\omega) - \frac{f(\omega)}{n}$$

and $\bar{f}(\omega) = \bar{f}(T\omega)$ a.e. since $f = \infty$. Thus $\bar{f}(T\omega) = \bar{f}(\omega)$ a.e., so $\bar{f}(\omega) = \bar{f}(T\omega)$ a.e., proving $\bar{f}$ almost invariant.

If $B \in \mathcal{I}$ and $C = \{\omega : \hat{f}(\omega) \in B\}$, then $T^{-1}C = \hat{f}(T\omega) \in B$ whenever $\omega \in C$, so $TC \subset C$. Thus C is almost invariant, and the proof is complete.

8.3.19 Theorem. *If $f \in L^1$ and A is an almost invariant set of finite measure, then $\int_A f(\omega)\,d\omega = \int_A \hat{f}(\omega)\,d\omega$. Thus if μ is a probability space, $\hat{f} = E(f \mid \mathcal{I})$, where $\mathcal{I}$ is the σ-field of almost invariant sets.*

Proof. Restrict T and μ to the almost invariant set A. Since $\mu(A) < \infty$ we have $f^A \in L^1$ on A by 8.2.5, and $\int_A f^A\,d\mu = \int_A \hat{f}\,d\mu$. But

$$\int_A f^A\,d\mu = \int_A f\,d\mu \qquad \text{by 8.2.12.} \quad \square$$

In the ergodic case, $\hat{f}$ assumes a very special form.

8.3.20 Theorem. *If T is ergodic and $f \in L^1$, then $\hat{f}$ is constant a.e. If $\mu(\Omega) = \infty$ the constant is $c = 0$, if $\mu(\Omega) < \infty$, we have*

$$c = \frac{1}{\mu(\Omega)} \int_\Omega f\,d\mu$$

Thus on a probability space $\hat{f} = E(f)$ a.e.

Proof. By 8.3.8, $\hat{f}$ is almost invariant, so by 8.2.7(b) $\hat{f} = c$ a.e. If $\mu(\Omega) = \infty$ then $c = 0$ because $\int_\Omega \hat{f}\,d\mu = \int_\Omega f\,d\mu < \infty$ by 8.3.19. If $\mu(\Omega) < \infty$, then $c = \mu(\Omega)^{-1} \int_\Omega f\,d\mu$ by 8.3.19 (with $A = \Omega$). $\square$

If T is ergodic and μ is finite, consider the case $f = 1_A$. Then by 8.3.20 $\hat{f} = \mu(A)/\mu(\Omega)$ a.e., so $\lim_n n^{-1} \sum_{k=0}^{n-1} 1_A(T^k\omega)$, the relative frequency of visits to A, converges a.e. to the relative size of A in the probability of A if $\mu(\Omega) = 1$. Apparently we have a version of the strong law of large numbers, and in fact the pointwise ergodic theorem can be regarded as a generalization of this result. Let T be the measure-preserving transformation in 8.1.3 Example 4 and 8.2.8 Example 4, and recall the construction preceding it. If X is an integrable random variable on $(\Omega, \mathcal{F}, P)$, then

$$X^{(1)}(\omega) = d^{-1}[E(X(\omega, \omega_2, \ldots) \mid \mathcal{Z}(\omega_2, \omega_3, \ldots)) + \cdots + E(X(\omega, \omega_2, \ldots) \mid \mathcal{Z}(\omega_d, \omega_{d+1}, \ldots))],$$

in other words,

$$E(X^{(1)} \mid \mathcal{Z}) = d^{-1}[E(X(\omega_1, \omega_2, \ldots) \mid \mathcal{Z}(\omega_2, \omega_3, \ldots)) + \cdots + E(X(\omega_d, \omega_{d+1}, \ldots) \mid \mathcal{Z}(\omega_{d+1}, \omega_{d+2}, \ldots))].$$

By 9.2.9,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \to E(f \mid \mathcal{I})(x)$$

where $\mathcal{I}$ is the σ-field of almost invariant sets. If P is ergodic, the limit is $E(f)$ by 9.2.10. In particular, let $X_0, X_1, \ldots$ be identically distributed random variables with finite expectation. If $x = (x_0, x_1, \ldots)$ and T is the shift, $x = X(\omega)$ where

$$\omega^{-1}(H) = \{\omega : (X_0(\omega), X_1(\omega), \ldots) \in H\}, \quad H \in \mathcal{R}^\infty,$$

is the content of the strong law of large numbers (see 6.1.5).

In the next section, it will be necessary to consider probability measures P and P_T that are both preserved by a fixed measurable transformation T on $(\Omega, \mathcal{F})$. If $P_T \ll P$ and P is ergodic (relative to T), then $P_T = P$, i.e., if $P_T \neq P$ and P is ergodic, they must be identical or mutually singular.

9.3.11 *Theorem.* Let P and P_T be ergodic probability measures on $\mathcal{F}$. Then either $P = P_T$ or $P \perp P_T$.

Proof. Suppose that $P_T(A) \neq P(A)$ for some $A \in \mathcal{F}$, and let

$$D = \omega : \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} I_A(T^k \omega) \neq P(A)\}, \quad i = 1, 2.$$

By 9.3.9 and 9.3.10, $P_i(D_i) = 1$ and $D_1 \cap D_2 = \emptyset$, so that $P_1 \perp P_2$.

Theorem 9.3.11 gives us a criterion for ergodicity (unfortunately, impractical).

9.3.12 *Theorem.* The probability measure P is ergodic relative to T iff the only probability measures P_T preserved by T such that P_T is absolutely continuous with respect to P are identical to P.

Proof. If $P_T \neq P$, then $P \perp P_T$, so P is not ergodic. If A is an invariant set with $P(A) = 0$ or $P(A^c) = 0$, by ergodicity, hence $P_T(A) = 0$ or $P_T(A^c) = 0$ by absolute continuity. By 9.3.11, $P_T \perp P$. But then P_T is both absolutely continuous and singular with respect to P, hence P_T is the zero measure, a contradiction.

Conversely, if P is not ergodic, let A be a T-invariant set with $0 < P(A) < 1$. Define $P_T(B) = P(B \cap A)/P(A)$. For $B \in \mathcal{F}$, $T^{-1}(B) \in \mathcal{F}$, so $P_T(T^{-1}(B)) = P(T^{-1}(B) \cap A)/P(A) = P(T^{-1}(B) \cap T^{-1}(A))/P(A) = P(A \cap B)/P(A) = P_T(B)$, so P_T is preserved by T. Now

$$\begin{aligned} P(T^{-1}E) &= \frac{P(T^{-1}E \cap A)}{P(A)} \\ &= \frac{P(T^{-1}E \cap T^{-1}A)}{P(A)} \qquad \text{since } T \text{ is invariant} \\ &= \frac{(vP)(A)}{P(A)} \qquad \text{where } v \text{ preserves } P \\ &= P(E). \end{aligned}$$

Thus P_1 is preserved by T.

Exercises

1. Let $f(x) = x \bmod 1$ for all real x. Recall Borel sets and Lebesgue measure. If j is the inclusion of $[0,1)$ into $\mathbb{R}$, show that $f^{(n)}$ does not converge in L^1 to j.

2. Let $\Omega = \mathbb{R}^{\mathbb{N}}$, $\mathcal{F} = [\mathcal{B}(\mathbb{R})]^{\mathbb{N}}$, $X_n(\omega) = \omega_n$, and take the unique probability measure making the X_n independent with $P\{X_n = 0\} = P\{X_n = 1\} = \frac{1}{2}$. (This is the usual model for an infinite sequence of Bernoulli trials with probability $\frac{1}{2}$ of success on a given trial.) If E is the one-sided shift, so that $T(\omega) = \omega'$, show that $\frac{1}{n}\sum_{i=1}^{n} f(T^i\omega)$ does not converge in L^1 to $\int f(\omega)\, P(d\omega) = \frac{1}{2}$.

3. Let T be a measure preserving transformation on $(\Omega, \mathcal{F}, P)$. If T is ergodic we know that for every $f \in L^1$, $f^{(n)}$ converges a.e. to a constant. Conversely, if $f^{(n)}(\omega)$ converges a.e. and if for every $f \in L^1$ there is a constant $c = c(f)$ such that $f^{(n)} \to c$ a.e., show that T is ergodic. Give a counter example to this statement if $P(\Omega) = \infty$.

4. Let T be an ergodic measure-preserving transformation on $(\Omega, \mathcal{F}, P)$ with $P(\Omega) < \infty$. Let f be a real valued Borel measurable function such that $\int f \, dP$ exists. If $\omega \in \Omega$, we know that $f^{(n)}$ converges a.e. to $[P(\Omega)]^{-1} \int_\Omega f \, dP$. Conversely, if $f^{(n)}$ converges a.e. to a finite limit, show that $f \in L^1$. (A special case of this result was considered in Ch. 2, Problem 1. Note also that the result fails when $P(\Omega) = \infty$, take $f \equiv 1$.)

5. (Riesz–Kakutani Theorem in a Hilbert Space.) Let U be a bounded linear operator on the Hilbert space H, and let U^* be the adjoint of U, characterized by the requirement that $(Uf, g) = (f, U^*g)$ for all $f, g \in H$. If $(Uf, Ug) = (f, g)$ for all $f, g \in H$ (i.e., U is a continuous linear transformation on H which preserves inner products), show that $UU^* = I$. (Hint: Given $h \in H$, there is a unique element $k \in H$ such that $(Uf, g) = (f, h)$ for every $f \in H$. Of course k depends on g and we may write $k = U^*g$. It follows from the basic properties of the inner product that U^* is a bounded linear operator on H.)

Exercises the following results

(a) The following conditions are equivalent and define a unitary operator. (That is, an invertible isometry.)
   i. $U^*U = UU^* = I$, the identity operator on H.
   ii. U is one-to-one onto, and $\langle Ux, Uy\rangle = \langle x, y\rangle$ for all $x, y \in H$.
   iii. U is onto and norm, and $\|Ux\| = \|x\|$ for all $x \in H$.

(b) The following conditions are equivalent and define an isometry.
   i. $U^*U = I$.
   ii. $\langle x, y\rangle = \langle Ux, Uy\rangle$ for all $x, y \in H$.
   iii. $\|x\| = \|Ux\|$ for all $x \in H$.

   For the remainder of the problem, U is an isometry on H.

(c) If $f \in H$, then $Pf = f$ iff $U^n f = f$ for all n.

(d) Define $A_n = \frac{1}{n}(I + U + U^2 + \cdots + U^{n-1})$, note that

$$A_n = \frac{1}{n}\sum_{k=0}^{n-1} U^k.$$

   If $S = \{f \in H : \lim_n A_n f \text{ exists in } H\}$, S is a closed subspace of H.

(e) Let $N = \{g - Ug : g \in H\}$. (The set of differences of H and its translate under U. That is, $N = \{f \in H : Uf = f\}$; note that N is a closed subspace of H, by continuity of U.) Let $N_\perp = \{y : \langle y, g\rangle = 0 \text{ for all } g \in N\}$. If we define $\tilde{f} = \lim_n A_n f$ (where the limit is taken in H), then

$$f \in N \quad \text{implies} \quad f \in S \quad \text{and} \quad \tilde{f} = f;$$
$$f \in N_\perp \quad \text{implies} \quad f \in S \quad \text{and} \quad \tilde{f} = 0.$$

(f) If $H = N \oplus N_\perp$, the closure of the subspace S is the orthogonal direct sum of N and $N_\perp$. (See the discussion after 3.2.11.)

(g) (Wiener-Ergodic Theorem) Let U be an isometry of an inner product space H, and let P be the projection of H on the space S of all elements invariant under U. Then, for every $f \in H$,

$$\lim_n \frac{1}{n}\sum_{k=0}^{n-1} \|U^k f - Pf\| = 0.$$

   If T is a measure-preserving transformation and U_T is the induced isometry on $L^2(\Omega, \mathcal{F}, \mu)$, then $\tilde{f}$ is the conditional expectation of f with respect to the σ-field $\mathcal{S}$ of T-invariant sets.

(h) In addition, if T is ergodic and $\int f \, d\mu = 0$, then U_T is a unitary operator and $\|U_T^n - P\| = 1$ for all n.

6 Let T be a measurable transformation on $(\Omega, \mathcal{F}, P)$. Consider the linear space of finite signed measures on $(\Omega, \mathcal{F})$ that $\mathcal{N}$ is the convex set of probability

measures preserved by T. Show that T is ergodic relative to $\mathcal{E}$ iff P is an extreme point of K, that is, P cannot be expressed as $\lambda P_1 + (1-\lambda)P_2$ with $0 < \lambda < 1$, $P_1 \in K$, $P_2 \in K$, $P \neq P_1$, $P \neq P_2$.

5.3. This problem gives many (but not quite all) equivalent formulations of ergodicity. In particular, some of the conditions involve convergence in probability rather than almost sure or L^1 convergence.

Let T be a measure-preserving transformation on the probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{E}$ be a field of sets whose minimal monotone class is $\mathcal{F}$. Show that the following conditions are equivalent:

(a) T is ergodic.

(b) $P(A) = P(A)^2$ or $P(A) = 0$ for each $A \in \mathcal{E}$, where

$$P(A) = \lim_{n} \frac{1}{n} \sum_{i=1}^{n} P(A \cap T^{-i}A)$$

(c) $P(A) = P(A)^2$ or $P(A) = 0$ for each $A \in \mathcal{F}$.

(d) $P(A) = P(A)^2$ or $P(A) = 0$ for each $A \in \mathcal{F}_0$.

(e) $P(A) = P(A)^2$ or $P(A) = 0$ for each $A \in \mathcal{F}_1$.

(f) $n^{-1} \sum_{i=1}^{n} P(A \cap T^{-i}B) \to P(A)P(B)$ for all $A, B \in \mathcal{F}_0$.

(g) $n^{-1} \sum_{i=1}^{n} |P(A \cap T^{-i}B) - P(A)P(B)| \to 0$ for all $A, B \in \mathcal{F}_0$.

If T is a one-sided shift transformation, we may take $\mathcal{F}_0$ to be the field of measurable cylinders. Further, since if the coordinate random variables take on only finitely many possible values, a measurable cylinder is a finite disjoint union of sets of the form $\{X_1 = a_1, \ldots, X_n = a_n\}$. The condition (f) is equivalent to the following statement:

Given each $n = (a_1, \ldots, a_n)$, $a_i = 1, \ldots, k$, with each a_i belonging to the coordinate space, let N_n be the number of times that $a_1, \ldots, a_n$ occur in succession in the first $n+n-1$ coordinates, that is,

$$N_n(\omega) = \sum_{i=1}^{n} f_n(T^i\omega), \qquad \text{where} \qquad f_n = 1_{\{X_1 = a_1, \ldots, X_n = a_n\}},$$

then $n^{-1}N_n$ converges in probability to $p(a_1) \cdots p(a_n)$.

Note also that by the ergodic theorem, given any one-sided shift with coordinate random variables X_i, there is a two-sided shift with coordinate random variables Y_i such that $(X_1, X_2, \ldots)$ and $(Y_1, Y_2, \ldots)$ have the same distribution. For example, specify that $(Y_1, Y_2, \ldots)$

Thus the one-dimensional distributions of $(X_0, X_1, X_2, \ldots)$ and $(X_1, X_2, X_3, \ldots)$ are identical. Similarly, the two-dimensional distributions of $(X_0, X_1, X_2, \ldots)$ and $(X_1, X_2, X_3, \ldots)$ are identical. Continuing in this way, we see that the two processes have identical finite-dimensional distributions. Thus $(X_0, X_1, X_2, \ldots)$ and $(X_1, X_2, X_3, \ldots)$ have the same distribution on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$. Corollary 8.2.1 shows that the two-sided shift is ergodic if and only if the corresponding one-sided shift is ergodic; by 8.2.1, the second statement of Kolmogorov's zero-one law implies the first.

8.4 Applications to Markov Chains

If T is a two-sided shift on $\mathbb{R}^{\mathbb{Z}}$ and the coordinate random variables X_n are independent, we have seen that T is ergodic. We also say that the sequence (X_n) is ergodic. In this section we obtain a large class of processes in which T is ergodic, but the X_n are not independent.

First we must consider some general properties of strictly stationary processes. If $X_0, X_1, \ldots$ are the coordinate random variables on a two-sided shift, recall from 8.1.2 that the σ-field $\mathcal{I}$ of invariant sets is contained in $\mathcal{F}_n = \bigcap_{m \geq n} \mathcal{F}_m^n$, where $\mathcal{F}_m^n = \mathcal{F}(X_m, X_{m+1}, \ldots)$. Intuitively, the smaller the σ-field $\mathcal{F}_n$, the more symmetry we have, because $\mathcal{F}_n$ is measurable with respect to the coordinates X_m for all $m \geq n$. We prove that the σ-field of almost invariant sets is essentially contained in $\mathcal{F}_n$.

8.4.1 Theorem *If A is an almost invariant set, there is a set $B \in \mathcal{F}_n$ such that $P(A \Delta B) = 0$.*

Proof. By 8.2.2(iii), there is a strictly invariant set B with $P(A \Delta B) = 0$. Now if $C \in \mathcal{F}$ then $T^{-n}(C) \in \mathcal{F}_n$; for if $C = \{(x_0, x_1, \ldots) : \ldots\} = C$, then $T^{-n}(C) = \{(x_n, x_{n+1}, \ldots) : \ldots\} \in \mathcal{F}_n$. (Actually, $C = T^{-n}(C)$ here since the X_k are coordinate random variables.) But $B = T^{-n}(B)$, so $B \in \mathcal{F}_n$ for all n, hence $B \in \mathcal{F}_\infty$. ∎

The proof of 8.4.1 shows that all strictly invariant sets belong to $\mathcal{F}_\infty$; however, an almost invariant set might not belong to $\mathcal{F}_1$. For example, let $A = \{\omega : X_0(\omega) = Y_0(\omega) \text{ for all } \omega\}$; then $A = T^{-1}A$, so that A is almost invariant by 8.2.1, but $A \notin \mathcal{F}_\infty$.

The following fact about conditional probabilities will be used.

8.4.2 Lemma *If T is a measure-preserving transformation on $(\Omega, \mathcal{F}, P)$, and $\mathcal{G}$ is a sub-σ-field of $\mathcal{F}$ then for any $A \in \mathcal{F}$,*

$$P(A \mid \mathcal{G})(T\omega) = P(T^{-1}A \mid T^{-1}\mathcal{G})(\omega) \quad \text{a.e.}$$

In particular, if T is a shift transformation, then

$$P(A \mid X_1, X_2, \ldots)(T\omega) = P(T^{-1}A \mid X_0, X_1, \ldots)(\omega) \quad \text{a.e.}$$

Proof. If $T^{-1}\mathcal{G} \subset \mathcal{G} \subset \mathcal{F}$, then

$$\int_{T^{-1}G} P(T^{-1}A \mid \mathcal{G})(T\omega)\, dP = \int_{T^{-1}G} P(T^{-1}A \mid T^{-1}\mathcal{G})\, dP$$

But since T is measure preserving,

$$\int_{T^{-1}A} f(Tx)\,dP(x) = \int_A f(x)\,dP(x)$$

$$= \int_{x_1=0}^{\infty} F(x_1 \mid x_2 = y)\,dP(x_2 \le y) \qquad \text{by (6.2)}$$

Since $F(x_1 \mid x_2)(Tx)$ is $T^{-1}\mathcal{B}$-measurable, the result follows. □

We may give an intuitive interpretation of (6.2). If T is a shift transformation on $\Omega = \{\omega : \omega = (\omega_1, \omega_2, \ldots), \omega_i \in \mathbb{R}^1\}$, and $P(T^{-1}A) = P(A)$, then $P(A \mid X_2)(Tx)$ is the probability that $Tx \in A$, given that the 2nd coordinate of Tx is $X_2(Tx) = x_3$. $P(A \mid X_2)(x)$ is the probability that $x \in A$, given that the 2nd coordinate of x is $X_2(x) = x_2$. These two expressions agree.

Our interest is in sequences having the Markov property, defined as follows.

6.6.3 *Definition.* A sequence of random variables $\{X_n : n \ge 0\}$ is said to have the *Markov property* if for each $B \in \mathcal{F}(X_{n-1}, X_{n-2}, \ldots)$, $n = 0, 1, \ldots$,

$$P(B \mid X_n, X_{n-1}, \ldots, X_0) = P(B \mid X_n) \qquad \text{a.e.}$$

If $\{X_n\}$ is a Markov chain, $n \ge 0$, $A \in \mathcal{B}$, $0 \le n < j$, then

$$P(X_{n+1} \in A, \ldots, X_{j-1} \in A, X_j \in B \mid X_0 = x_0, \ldots, X_n = x_n)$$
$$= P(X_{n+1} \in A, \ldots, X_{j-1} \in A, X_j \in B \mid X_n = x_n)$$

for all $x_0, \ldots, x_{n-1}$ in the state space. This is the Markov property holds when B is a measurable cylinder, hence for all $B \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ by the monotone class theorem.

In the Markov case, almost invariant sets assume a special form.

6.6.4 *Theorem.* For the one-sided stationary random variables X_n of a measure-preserving transformation, the Markov property holds, and $\mathcal{I}$ is an almost invariant set, there is a set $B \in \mathcal{B}^\infty$ of the form $B = C \times C \times \cdots$, $C \in \mathcal{B}$, such that $A = B$ a.e., $A \in \mathcal{I}$, and $P(A \cap B) = 0$.

Proof. By 6.6.1 there is a set B of the full form $B = C \times C \times \cdots$ such that $P(A \cap B) = 0$. Since $B \in \mathcal{F}(X_0, X_1, \ldots)$, we have $P(B \mid X_0, X_1, \ldots, X_n) = P(B \mid X_n)$. But $B \in \mathcal{F}(X_n)$, so $P(B \mid X_0, X_1, \ldots, X_n) = P(B \mid X_n)$. Now

$$P(A \mid X_0, \ldots, X_n) = P(A \mid X_0, X_1, \ldots)$$

Since $A \in \mathcal{F}(X_1, X_2, \ldots)$ we have

$$P(A \mid X_1, X_2, \ldots) = P(A \mid X_2, X_3, \ldots) = I_A \quad \text{a.s.}$$

since $A \in \sigma(X_2, X_3, \ldots)$. Thus $P(A \mid X_0) = I_A$ a.s.

Conversely, if $A \in \mathcal{F}(X_0, X_1, \ldots)$ then $P(A \mid X_0) = I_A$ a.s. Thus $P(A^c) = 0$ since $P(A \mid X_0) = I_A$ a.s. hence in particular $A \in \mathcal{F}$. Now

$$T^{-1}(A) = \{\omega : P(A \mid X_1)(T\omega) = I_A(T\omega)\} \supseteq A$$

$$= \{\omega : P(A \mid X_0, X_1)(\omega) = I_A(\omega)\} \supseteq A \quad \text{a.s.}$$

by 8.4.2 and the almost invariance of A. Thus $T^{-1}(A) = A$ a.s. Since T preserves P, $P(A) = P(T^{-1}(A)) = P(A \cap A) = P(A)^2$ so $P(A) = 0$ or 1. Since $P(A^c) = 0$, we must have $P(A) = 1$, so $P(A \mid X_0) = I_A$ a.s. for all $A \in \mathcal{F}$, in particular, $P = P(A \mid X_0)$ a.s.

Now $P(A \mid X_0)$ is $\sigma(X_0)$ measurable, hence can be represented as $g(X_0)$ for some Borel measurable $g : S \to [0, 1]$ (see 5.1.1). Thus we have

$$\omega \in A \quad \text{iff} \quad I_A(\omega) = 1$$

$$\text{iff} \quad P(A \mid X_0)(\omega) = 1$$

$$\text{iff} \quad g(X_0(\omega)) = 1 \quad \text{where} \quad g(x) = P(A \mid X_0 = x)$$

8.4.5 Corollary. Under the hypothesis of 8.4.2, a set $A \in \mathcal{F}$ is almost invariant iff A is of the form $\{X_0 \in C\}$ for all n for some $C \in \mathcal{B}(S)$.

Proof. If A is almost invariant, by 8.4.1, A is of the form $\{X_1 \in C\}$. But $A = T^{-1}(A)$ also, so $\{X_1 \in C\} = \{X_2 \in C\}$ a.s. Inductively, $A = \{X_n \in C\}$ for all n. Conversely, every set of this form has $A = T^{-1}(A)$ hence is almost invariant by 8.4.1. ∎

We can now have the following criterion for ergodicity of a Markov sequence:

8.4.6 Theorem. If X_n has the Markov property then $\{X_n\}$ is not ergodic iff there is a set $C \in \mathcal{B}(S)$ such that $0 < P(X_n \in C) < 1$ for all $n \geq 1$.

Proof. Apply 8.4.5 and 8.3.2(ii). ∎

We now apply these results to Markov chains. If $\{X_n\}$ is a Markov chain, assume the initial distribution $\{\mu_i\}$ is a stationary distribution, so that $\{X_n\}$ is strictly stationary, and the machinery of ergodic theory is applicable. See

Acad. 1973, pp. 230–240, for the appropriate basic concepts and material on Markov chains.)

5.4.7 Theorem.

(a) If there is exactly one positive recurrent class C, then $\{X_n\}$ is ergodic.

(b) If there are at least two positive recurrent classes C_1 and C_2, and $\sum_{i,j \in C_1} p_{ij} > 0$, $\sum_{i,j \in C_2} p_{ij} > 0$, then $\{X_n\}$ is not ergodic.

Proof. (a) Since the stationary distribution assigns probability 1 to C, we may as well assume that C is the entire state space. Let D be a minimally proper subset of C, and let $\beta = \min_{i,j \in D} p_{ij} > 0$. By hypothesis, if n is large enough, then j must be visited from i with probability at least β, i.e.,

$$p_{ij}(n) \geq \beta \text{ for all } n \geq N = \sum_{r=1}^{d} (r-1)\lambda_r + N_0 \text{ for some } N_0 \text{ if } X_0 = i, \; X_n = j \in C = D.$$

By Exercise 5.4.6, $\{X_n\}$ is ergodic.

(b) It is impossible to start from a recurrent class since

$$P[X_n = i, \text{ for all } n] = P[X_0 = i] = \sum_{i \in C_1} \pi_i = 0, 1.$$

But since $\sum_{i \in C_1} \pi_i < 1$, $\{X_n\}$ is not ergodic. ■

The case in which there are no positive recurrent classes is not discussed in 5.4.7 because in this case, there is no stationary distribution for the chain. Note that if there is exactly one positive recurrent class, the stationary distribution is unique.

We now have many examples of ergodic sequences $\{X_n\}$, where the X_n need not be independent. For example, consider a finite Markov chain such that every state is reachable from every other state. The state space then forms a single equivalence class, necessarily recurrent positive. Thus if the initial distribution is the unique stationary distribution, the sequence $\{X_n\}$ is ergodic.

Now suppose $\{X_n\}$ is an ergodic Markov chain, and assume that the entire state space forms a positive recurrent class. Then by the pointwise ergodic theorem, if $g = h(X_0, X_1, \ldots)$, then

$$\frac{1}{n} \sum_{i=1}^{n} g(T^i \omega) \to E(g) \quad \text{a.e.}$$

We have seen in Example 8.3.3 that the initial distribution π is stationary, but this result holds regardless of the initial distribution. For

$$\begin{aligned} 0 &= P\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} f(T^m\omega) = \sum_{i=1}^{k} f(i)\pi_i\right\} \\ &= \sum_{i=1}^{k}\pi_i P_i\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} f(T^m\omega) = \sum_{i=1}^{k} f(i)\pi_i \mid X_1 = i\right\} \end{aligned}$$

Since $\pi_i > 0$ for all i, we have $P_i\{n^{-1}\sum_{m=1}^{n} f(T^m\omega) \to \sum_{i=1}^{k} f(i)\pi_i \mid X_1 = i\} = 0$ for all i, and the result follows.

If a Markov chain has exactly one positive recurrent class, and therefore a unique stationary distribution $\{\pi_i\}$, then the mean recurrence time of state j, that is, the average number of steps required to return to j when an initial state $\neq j$, is the reciprocal of the probability π_j that the chain will be in state j at any particular time. This is intuitively reasonable. If, say, $\pi_j = \frac{1}{10}$, then in the long run, we are in state j one time out of ten, and the average length between successive times in state j is 10.

We are going to prove a more general result of this type.

8.4.5 *Theorem.* Let T be a measure preserving transformation on the probability space $(\Omega, \mathcal{F}, P)$. If $A \in \mathcal{F}$, let

$$A_n = \{\omega \in A : T^m\omega \notin A, m = 1, \ldots, n-1, T^n\omega \in A\},$$

thus A_n is the set of points $\omega \in A$ such that $T^n\omega$ returns to A for the first time at $n = n$.

Define the recurrence time of A by $r_A(\omega) = n$ if $\omega \in A_n$, $n = 1, 2, \ldots$, $r_A(\omega) = \infty$ if $\omega \in A^c = \Omega - A$. Define T_A arbitrarily on A^c. Then

$$\int_A r_A(\omega)\, dP(\omega) = P\left(\bigcup_{n=1}^{\infty} T^{-n}A\right).$$

Before giving the proof, let us go over it informally. The idea is that the theorem is "obvious" if T is ergodic and $P(A) > 0$. If $B = \bigcup_{n=1}^{\infty} T^{-n}A$, then $TB \subset B$, so if T is ergodic, P is invariant by 8.2.4, and since $P(B) \geq P(A) > 0$, $P(B)$ must be 1. Thus if $Q(B) = P(B \mid A) = P(B \cap A)/P(A)$, $B \in \mathcal{F}$, we have

$$\int_A r_A(\omega)\, dQ(\omega) = \frac{1}{P(A)}$$

If P is not ergodic, then $A = \{x : r_A(x) < \infty\}$, then $\int_A r_A\, dP$ is the average length of time required for X_n to return to the set A, given that the initial point X_1

belongs to C. Thus the mean recurrence time of C is the reciprocal of the probability that the process will be in C at any particular time.

PROOF Let $C_n = S^{(n)}$, $B_n = C_n^c$, $n = 0, 1, \ldots$, take $C_0 = B_0$. Then, with increasing values of n, we have

$$\begin{aligned} P(A_{n-1}) &= P(C_0 \cdots C_{n-1}) \qquad n \geq 1 \\ &= P(C_{n-1} \mid C_0 \cdots C_{n-2}) P(C_0 \cdots C_{n-2}) \\ &= P(C_{n-1} \mid C_0 \cdots C_{n-2}) P(C_{n-2} \mid C_0 \cdots C_{n-3}) P(C_0 \cdots C_{n-3}) \\ &\quad \vdots \\ &= P(C_{n-1} \mid C_0 \cdots C_{n-2}) \cdots P(C_2 \mid C_0 C_1) P(C_1 \mid C_0) P(C_0) \end{aligned}$$

where $p_k = P(C_{n-1} \mid C_0 \cdots C_{n-2})$, $k \geq 1$. When $k = 0$ we have

$$\begin{aligned} P(A_1) &= P(C_0 C_1) \\ &= P(C_1 \mid C_0) P(C_0) \\ &= [1 - P(B_1 \mid C_0)] P(C_0) = P(C_0) - P(B_1 C_0) \\ &= 1 - 2P(B_1) + P(B_1 B_0), \end{aligned}$$

hence we have $P(A_{n-1}) = p_n - 2p_{n-1} + p_{n-2}$, $n \geq 2$, if we take $p_1 = 1$. Now

$$\begin{aligned} \int_0^\infty t \, dF &= \sum_{n=0}^{\infty} (n+1) P(A_{n-1}) \\ &= \lim_{N \to \infty} \sum_{n=0}^{N-1} (n+1)(p_n - 2p_{n-1} + p_{n-2}) = 2 \sum_{n=0}^{N-1} p_n - \sum_{n=0}^{N-1} p_{n-1} - (N-1) p_{N-1} \\ &= \lim_{N \to \infty} [1 + p_{N-1} - (N-1) p_{N-1}] = \lim_{N \to \infty} [1 + p_{N-1}(2 - N)]. \end{aligned}$$

Now $p_{N-1} = P(B_1 \cdots B_{N-2} C_{N-1}) \geq 0$, and

$$p_{N-1} \leq P\left(\bigcap_{n=1}^{N-1} B_n\right) = 1 - P\left(\bigcup_{n=1}^{N-1} C_n\right) \leq 1 - P\left(\bigcup_{n=1}^{N-1} S^{(n)}\right).$$

Since $1 - P(A_{n-1}) = P(A_{n-1}^c) = p_n$ has a limit and p_n approaches a finite limit, $(N-1)p_{N-1}$ must approach a finite limit. But if we set $B_1 \cdots B_{N-1} C_{N-1}$ are disjoint so that $\sum_{n=1}^{N-1} p_{n-1} \leq 1$, this implies that $\lim_{N \to \infty} (N-1) p_{N-1} = 0$. Thus $\int_0^\infty t \, dF = 1$.

Problems

1. The unpredictability of a transformation T is said to be a 0–1 law if the σ-algebra $\mathcal{I}$ of events invariant under T contains only events which have probability 0 or 1. If Kolmogorov's 0–1 law holds for a Kolmogorov shift in a space, show that in such a space every Kolmogorov shift is mixing.

2. Let P be the stochastic transformation associated with a finite Markov chain where the initial distribution is stationary. Assume that $q_i > 0$ for all i. If $X_1 = C_1$, then $P(X_n = i \text{ for some } n \geq 2 \mid X_1 = C_1)$ or $P(X_n = i \text{ for some } n \geq 2 \mid X_1 = C_2)$ or ... must as well be removed from the state space.

   Show that P is mixing iff $\lim_{n} p_{ij}^{(n)} = q_j$ independent of i as $n \to \infty$, where $\sum_j q_j = 1$. (Necessarily, $q_j = \pi_j$, and the stationary distribution is unique; see Ash, 1970, pp. 235–237.)

3. Let $X_1, X_2, \ldots$ be real random variables, and let $S_n = \sum_{k=1}^{n} X_k$, $n = 1, 2, \ldots$.

   (a) If $\{\omega : \omega \text{ is a limit point of the sequence } S_1(\omega), S_2(\omega), \ldots\}$ and $A = \{\omega : \omega \in \Omega : S_n(\omega) \in A \text{ i.o.}\}$, show that $\sigma^{-1}(A) = A$. (Hint: Express A_ε as $\bigcap_{r} \bigcup_{n} \{\omega : |S_n(\omega) - r| < \varepsilon\}$, where r ranges over the rationals.)

   (b) Let α be a fixed positive integer. If $N_n(\omega)$ is the number of distinct points in $\{S_n(\omega), S_{n+1}(\omega), \ldots, S_{n+\alpha-1}(\omega)\}$, $n = 1, 2, \ldots$. Use the ergodic theorem to show that

$$\limsup_{n \to \infty} \frac{N_n(\omega)}{n} \leq \frac{E(|X_1|)}{\alpha} \quad \text{a.e.}$$

   (c) Show that $\limsup_{n \to \infty} |S_n| = \infty$ a.e. or $S_n \to S(\omega)$ a.e.

   (d) Let N_α be the cardinality of the set $\{S_1, S_2 - S_1, S_3 - S_2, \ldots, S_\alpha - S_{\alpha-1}\}$ and $N_\alpha = 1$ if S_n is a proper sequence. Use the ergodic theorem to show that $n^{-1} N_\alpha(\omega) \to E(N_\alpha)$ a.e.

   (e) Show that $\liminf_{n \to \infty} |S_n| \geq E(|X_1|)$ a.e., and conclude that $|S_n| \to \infty$ a.e. or $|S_n| \to E(|X_1|)$ a.e.

8.3 The Shannon–McMillan Theorem

We now apply the pointwise ergodic theorem to prove a basic result of information theory. Consider a shift transformation with ergodic random

variables X_i taking values in a countable set. We define another unusual sequence of random variables on $(X_1, X_2, \ldots, X_n)$ as follows.

If $X_1(\omega) = x_1, X_2(\omega) = x_2, \ldots, X_n(\omega) = x_n$, we set

$$p(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = p(x_1, x_2) p(x_2, x_3) \cdots p(x_{n-1}, x_n).$$

For example, if $p(X_1 = 1) = p(X_1 = 2) = \frac{1}{2}$, $p(1,1) = p(2,1) = \frac{1}{4}$, $p(1,2) = \frac{1}{2}$, then $p(X_1(\omega))$ has the value $\frac{1}{2}$ with probability $\frac{1}{2}$ (namely, when $X_1 = 1$ or 2) and $p(X_2) = \frac{1}{4}$ with probability $\frac{3}{4}$ (that is, when $X_1 = 1$).

Similarly, define the random variable $p(X_1, X_2) p(X_2, X_3) \cdots p(X_{n-1}, X_n)$ by specifying that if $X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n$, then $p(X_1, X_2) \cdots p(X_{n-1}, X_n) = p(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$, assuming $p(X_1 = i_1, \ldots, X_n = i_n) > 0$.

The Shannon–McMillan theorem is essentially about the convergence of $-\frac{1}{n} \log p(X_1, \ldots, X_n)$ to a constant. If S is ergodic, we have convergence to a constant H almost everywhere in Ω. Before turning to the proof, let us look at the intuitive interpretation of the convergence statement. For $\varepsilon > 0$, we require only that $-\frac{1}{n} \log p(X_1, \ldots, X_n)$ converge to H in probability; that is traditional in information theory terminology, but we shall follow the practice here. Switching to natural logs involves only a multiplication constant.)

Given $\varepsilon > 0$, $n \geq 1$, let

$$A_n = \left\{ x : \left| -\frac{1}{n} \log p(X_1, \ldots, X_n) - H \right| \leq \varepsilon \right\}.$$

That is, $x \in A_n$ means $p(x_1, \ldots, x_n)$ is between $2^{-n(H+\varepsilon)}$ and $2^{-n(H-\varepsilon)}$. If $-\frac{1}{n} \log p(X_1, \ldots, X_n) > H$ is probability, then $P(A_n) \geq 1 - \varepsilon$ for sufficiently large n.

Now let S_n be the set of all sequences of length n with values in the coordinate space corresponding to points in A_n; that is, S_n is the set of all sequences $(X_1(\omega), \ldots, X_{n-1}(\omega))$, $\omega \in A_n$. If $(i_1, \ldots, i_{n-1}) \in S_n$, then $2^{-n(H+\varepsilon)} < P(X_1 = i_1, \ldots, X_n = i_n) < 2^{-n(H-\varepsilon)}$; furthermore, $P(X_1 = i_1, \ldots, X_n = i_n \mid (i_1, \ldots, i_{n-1}) \in S_n) = P(X_1 = i_1, \ldots, X_n = i_n) / P(A_n)$, so that $1 - \varepsilon < P(X_1 = i_1, \ldots, X_n = i_n) < 1$.

Thus each sequence in S_n has probability between $2^{-n(H+\varepsilon)}$ and $2^{-n(H-\varepsilon)}$, and the total probability assigned to S_n is between $1 - \varepsilon$ and 1. Consequently, the maximum number of sequences in S_n is $1 \cdot 2^{n(H+\varepsilon)} = 2^{n(H+\varepsilon)}$, and the minimum number is $(1 - \varepsilon) 2^{n(H-\varepsilon)} = (1 - \varepsilon) 2^{n(H-\varepsilon)}$.

Thus, roughly, we have shown that there are approximately 2^{nH} sequences of length n, each with probability approximately 2^{-nH}; the remaining sequences are

negligible, that is, have total probability as close to 1 as desired. In information theory, this is referred to as the asymptotic equipartition property. If $|X_0|$ represents the number of the "information symbols" in a language with a given vocabulary, then of all the $n^{|X_0|}$ possible sequences of length n, only 2^{nH} are reasonably to be expected to appear.

The number H will turn out to be the entropy of the sequence $\{X_n\}$, and we must discuss the concept before going any further.

8.5.2 *Definition.* If X is a discrete random variable (or random vector), define the entropy (also called the uncertainty) of X as

$$H(X) = -\sum_x p(x) \log p(x)$$

where $p(x) = P(X = x)$. If Y and X are discrete, define the conditional entropy of Y given $X = x$ as

$$H(Y \mid X = x) = -\sum_y p(y \mid x) \log p(y \mid x)$$

where $p(y \mid x) = P(Y = y \mid X = x)$, also define the conditional entropy of Y given X as a weighted average of the $H(Y \mid X = x)$, namely,

$$H(Y \mid X) = \sum_x p(x) H(Y \mid X = x)$$
$$= -\sum_{x,y} p(x, y) \log p(y \mid x).$$

The joint entropy of X and Y is defined by

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

since $H(X, Y)$ is the entropy of the random vector (X, Y), nothing new is involved.

Note that entropy is always nonnegative. If the random variables are discrete but have a countably infinite set of values, the entropy may or may not be obtained. Difficulties were overcome by defining $0 \log 0 = 0$ since $x \log x \to 0$ as $x \to 0$.

It is often convenient to express entropy in terms of the random variables $p(X)$ introduced at the beginning of this section, we have

$$H(X) = E[-\log p(X)], \qquad H(X \mid Y) = E[-\log p(X \mid Y)].$$

We now establish a few properties of entropy.

5.5.2 Theorem. Let X, Y, and Z be discrete random variables with finite entropy.

(a) If $p_1, p_2, \ldots$ and $q_1, q_2, \ldots$ are two discrete probability distributions with $\sum_i p_i = \sum_i q_i = 1$, and if $-\sum_i p_i \log p_i$ exists and $-\sum_i p_i \log q_i$ exists, then

$$-\sum_i p_i \log p_i \le -\sum_i p_i \log q_i$$

with equality if $p_i = q_i$ for all i.

(b) $H(X, Y) \le H(X) + H(Y)$, with equality iff X and Y are independent.

(c) $H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$.

(d) $H(f(X)) \le H(X)$, with equality iff f is one-to-one.

(e) If X takes on n possible values $x_1, \ldots, x_n$, then $H(X) \le \log n$, with equality iff $p(x_i) = 1/n$, $i = 1, \ldots, n$.

(f) $H(X, Z \mid Y) \le H(X \mid Y) + H(Z \mid Y)$, with equality iff X and Z are conditionally independent given Y, that is,

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \qquad \text{for all } x, y, z$$

(g) $H(X, Y, Z) + H(Y) \le H(X, Y) + H(Y, Z)$.

(h) $H(Z \mid X, Y) \le H(Z \mid X)$, with equality iff X and Z are conditionally independent given Y and X.

Proof. (a) First consider the natural logarithm. Since $\ln x = 1$ is the tangent to $\log x$ at $x = 1$, we have $\log x \le x - 1$, with equality iff $x = 1$. Thus $\log(q_i/p_i) \le q_i/p_i - 1$, with equality iff $q_i = p_i$. Hence, provided $p_i \ne 0$, $i = 1, 2, \ldots$,

$$p_1 \log q_1 - p_1 \log p_1 + p_2 \log q_2 - p_2 \log p_2 + \cdots \le q_1 - p_1 + q_2 - p_2 + \cdots$$

with equality iff $q_i = p_i$. Summing over i yields the desired result. (Note that if $\sum_i p_i \log p_i = \sum_i p_i \log q_i$, then the sum is finite by hypothesis, so that $p_i = q_i$ for all i.)

(b) By (a), $-\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j) \le -\sum_{i,j} p(x_i, y_j) \log p(x_i)p(y_j)$, with equality iff $p(x_i, y_j) = p(x_i)p(y_j)$ for all i, j.

(c) This follows from the identity $p(x,y) = p(x)p(y \mid x) = p(y)p(x \mid y)$.
(d) This is immediate from (c) and (a).
(e) Apply (a) with $p_x = p(x)$ and $q_x = 1/n$.
(f) Since $H(X \mid Y = y) \le \log |\mathcal{X}|$ for each y such that $p(y) > 0$, and similarly for $H(X \mid Y = y)$, the proof of (f) shows that

$$H(X \mid Y) - H(X \mid Y = y) \le \log |\mathcal{X}| - H(X \mid Y = y),$$

with equality iff $p(x \mid y) = p(x)$ or $p(x) = 0$ or $p(y \mid x) = p(y)$ for all x, y. Multiply by $p(y)$ and sum over y to complete the proof.
(g) This follows since $p(x,z \mid y) = p(x \mid y)p(z \mid y)$.
(h) Apply (f) and (g).

We also need the following concept for stationary sequences.

8.5.3 Definitions and Comments. Let $X_1, X_2, \ldots$ be a stationary sequence of discrete random variables, and assume that $H(X_1) < \infty$. Define the entropy of the sequence as

$$H(\mathcal{X}) = \lim_{n \to \infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1).$$

Now

$$\begin{aligned} H(X_{n+1} \mid X_n, \ldots, X_2) &\le H(X_{n+1} \mid X_n, \ldots, X_1) && \text{by (8.5.2(b))} \\ &= H(X_n \mid X_{n-1}, \ldots, X_1) && \text{by stationarity,} \end{aligned}$$

and by (8.5.2(a)), $H(X_n \mid X_{n-1}, \ldots, X_1) \le H(X_{n-1}) = H(X_1) < \infty$. Thus the limit defining $H(\mathcal{X})$ exists and is finite.

(b) To apply (8.5.2(c)), it must be verified that $H(X_1, \ldots, X_n) < \infty$, but the proof of (8.5.2(b)) shows more: if $H(X_1) < \infty$, then

$$H(X_1, \ldots, X_n) \le H(X_1) + \cdots + H(X_n),$$

with equality iff $X_1, \ldots, X_n$ are independent. In the present case, $H(X_i) = H(X_1) < \infty$ for all i.

The entropy of $\{X_n\}$ may also be expressed as

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n).$$

To see this, we first prove by induction on n that (8.5.2(b)),

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1) \\ &\quad + \cdots + H(X_n \mid X_{n-1}, \ldots, X_1). \end{aligned}$$

Thus $-\log p(X_0, X_1, \ldots, X_{n-1})$ is the uniformly decreasing sequence of successive entropies in $H(X_0)$ and hence converges to $H(\mathcal{X})$.

We now begin the development of the Shannon–McMillan theorem. It will be convenient to consider a two-sided shift transformation T with coordinate random variables X_n, where the X_n take values in a countable set. It is assumed throughout that $H(X_0) < \infty$.

Mutual information will be significant, as the following result suggests.

5.5.4 Theorem. Let $Y = -\log p(X_0)$, $Y_n = -\log p(X_0 \mid X_{-1}, \ldots, X_{-n})$, $n > 0$. Then $\{Y, Y_1, Y_2, \ldots\}$ is a nonnegative submartingale (hence uniformly integrable) with respect to an increasing sequence of σ-fields $\mathcal{F}$.

Proof. Since $E(Y_n) = H(X_0 \mid X_{-1}, \ldots, X_{-n}) \le H(X_0)$, the Y_n are integrable. Now since the random variables are discrete, we may write

$$\begin{aligned} E(Y_n \mid & X_1 = x_1, \ldots, X_n = x_n) \\ &= -\sum_{x_0} p(x_0 \mid x_1, \ldots, x_n) \log p(x_0 \mid x_1, \ldots, x_n) \\ &= -\sum_{x_0} \frac{p(x_0, \ldots, x_n)}{p(x_1, \ldots, x_n)} \log \frac{p(x_0, \ldots, x_n)}{p(x_1, \ldots, x_n)}. \end{aligned}$$

This is of the form $\sum_i a_i \log a_i$, where $a_i \ge 0$, $\sum_i a_i = 1$, and hence is less than or equal to $\log |\sum_i a_i|$ by convexity. Thus

$$\begin{aligned} E(Y_{n+1} \mid & X_2 = x_1, \ldots, X_{n+1} = x_n) \\ &\le \log \left| \sum_{x_0} \frac{p(x_0, \ldots, x_n)}{p(x_1, \ldots, x_n)} \right| \\ &= \log 1 = 0 \\ &= E(Y_1 \mid X_1 = x_1, \ldots, X_n = x_n). \end{aligned}$$

Thus $E(Y_{n+1} \mid X_2, \ldots, X_{n+1}) \le E(Y_1 \mid X_1, \ldots, X_n)$, and since Y_1 is measurable relative to $\sigma(X_1, \ldots, X_n)$ the result follows. $\square$

We now show that the random variables Y_n are uniformly integrable. It will be convenient to assume that the conditional space is a subset of the positive integers. This assumption only avoids labeling, and can be done without changing the basic nature of $(Y_0, Y_1, \ldots)$.

8.5.5 Lemma. If c is any fixed positive integer, the random variables $W_n = Y_n I(|Y_n| \le c) - E[Y_n I(|Y_n| \le c) \mid \mathcal{F}_{n-1}]$, $n = 0, 1, \ldots$, are uniformly integrable.

Proof. For any positive integer k,

$$\int_{[|W_n| \ge k]} |W_n| \, dP = \sum_{i=k}^{\infty} \int_{[i \le |W_n| < i+1]} |W_n| \, dP$$

$$\le \sum_{i=k}^{\infty} (i+1) \, P[|W_n| \ge i]$$

But $|W_n| \ge i$ implies $|Y_n| \ge i$ and it follows that $P[|W_n| \ge i] \le 2 \, P[|Y_n| \ge i]$. Thus

$$P[|W_n| \ge i] = \sum_{\substack{|x_{n-1}| \le c, \ldots, |x_0| \le c \\ |x_n| \ge i}} p(x_n, \ldots, x_0)$$

$$\le \left| \sum_{x_{n-1}} \cdots \sum_{x_0} p(x_n, \ldots, x_0) \right|$$

$$\le 2 \, P[|Y_n| \ge i]$$

Consequently,

$$\int_{[|W_n| \ge k]} |W_n| \, dP \le 2 \sum_{i=k}^{\infty} (i+1) P[|Y_n| \ge i] \to 0$$

as $k \to \infty$, uniformly in $n \ge 1$.

8.5.6 Theorem. The nonnegative random variables $Z_0, Z_1, \ldots$ are uniformly integrable. Thus in 8.5.4 we also have $E_n \to E_\infty$ in L^1.

Proof. For any positive integer k,

$$\int_{[Z_n \ge k]} Z_n \, dP = \int_{[Z_n \ge k, Z_{n+1} \ge k]} Z_n \, dP + \int_{[Z_n \ge k > Z_{n+1}]} Z_n \, dP$$

$$\le \int_{[Z_n \ge k]} Z_{n+1} \, dP + \int_{[Z_n \ge k > Z_{n+1}]} Z_n \, dP$$

By 8.5.5, the second integral on the right approaches 0 as $k \to \infty$, uniformly in n. Now

$$\int_{\{N=k\}} g_n\,dP = \sum_{x_0 \in \Gamma} \left[\sum_{x_{-1},\ldots,x_{-n}} p(x_0,\ldots,x_{-n}) \cdot \log p(x_0 \mid x_{-1},\ldots,x_{-n}) \right]$$

$$= -\sum_{x_{-1},\ldots,x_{-n}} p(x_{-1},\ldots,x_{-n}) \cdot \sum_{x_0 \in \Gamma} p(x_0 \mid x_{-1},\ldots,x_{-n}) \log p(x_0 \mid x_{-1},\ldots,x_{-n}).$$

Write $\log p(x_0 \mid x_{-1},\ldots,x_{-n}) = \log p(x_0) + \log[p(x_0 \mid x_{-1},\ldots,x_{-n})/p(x_0)]$; the contribution due to $\log p(x_0)$ is

$$\sum_{x_0 \in \Gamma} \left[\sum_{x_{-1},\ldots,x_{-n}} p(x_0,\ldots,x_{-n}) \log p(x_0) \right] = \sum_{x_0 \in \Gamma} p(x_0) \log p(x_0)$$

and the remaining contribution is

$$\sum_{x_{-1},\ldots,x_{-n}} p(x_{-1},\ldots,x_{-n}) \sum_{x_0 \in \Gamma} p(x_0 \mid x_{-1},\ldots,x_{-n}) \log \frac{p(x_0)}{p(x_0 \mid x_{-1},\ldots,x_{-n})}.$$

An upper bound to this expression, obtained by switching to natural logs for convenience and using $\log_e x \le x - 1$, is

$$\sum_{x_0 \in \Gamma} p(x_0) - \sum_{x_0 \in \Gamma} p(x_0) = 0.$$

Therefore,

$$\int_{\{N=k\}} g_n\,dP \le -\sum_{x_0 \in \Gamma} p(x_0) \log p(x_0) \to 0$$

as $k \to \infty$, uniformly in n, since $H(X_0) < \infty$.

If $\varepsilon > 0$ is given, we may choose k fixed in such that $\int_{\{N > k\}} g_n\,dP < \varepsilon/2$ for all n. Then for sufficiently large λ, $\int_{\{g^* > \lambda\}} W_k\,dP < \varepsilon/2$ for all n, hence $\int_{\{g^* > \lambda\}} g_n\,dP < \varepsilon$ for all n. □

The other basic property of the g_n that we need is that $\sup_n g_n$ is integrable. We prove this after a preliminary lemma.

8.5.7 **Lemma.** If τ is a stopping time, define $\mathcal{F}_\tau = \{A : A \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n\}$. Let τ be a stopping time relative to $X_0, X_1, \ldots$. Then

$$p_q^{(n)} = -\log P(X_\tau = j \mid X_{\tau-1} = i, \ldots, X_0 = i_{\tau-1}, \tau = n)$$

is $\mathcal{F}_\tau$-measurable. If $\tau < \infty$ a.s., then

$$P_{ij}(\omega) = \left\{ \begin{array}{ll} \inf\{n : X_n = j\} & \text{if } \omega \in A_i \\ 0 & \text{otherwise} \end{array} \right\}, \qquad E_q^{(n)}(\omega) = \left\{ \begin{array}{ll} \inf\{n : p_q^{(n)} < \infty\} & \text{if } \omega \in A_i \\ 0 & \text{otherwise} \end{array} \right\},$$

are $\mathcal{F}_\tau$-measurable.

Proof. Clearly $\{\omega : p_q^{(n)}(\omega) < \infty\} \in \mathcal{F}_\tau$. Hence $E_q^{(n)}(\omega) = n$ on $A_i \cap \{\omega : p_q^{(n)}(\omega) < \infty\} \in \mathcal{F}_\tau$. Now since $E_q^{(n)}(\omega)$ is finitely-valued, the σ-field $\mathcal{F}(E_q^{(n)}, X_0, \ldots, X_{n-1})$ is

$$\begin{aligned} \mathcal{F}(E_q^{(n)}, X_0, \ldots, X_{n-1}) &= \bigcup_{i_0, \ldots, i_{n-1}} A_{i_0} \cap \cdots \cap A_{i_{n-1}} \cap \{\omega : E_q^{(n)}(\omega) = n\} \\ &= \bigcup_{i_0, \ldots, i_{n-1}} A_{i_0} \cap \cdots \cap A_{i_{n-1}} \cap \{\omega : \tau(\omega) = n, X_0(\omega) = i_0, \ldots, X_{n-1}(\omega) = i_{n-1}\} \\ &= \bigcup_{i_0, \ldots, i_{n-1}} \{\omega : \tau(\omega) = n, X_0(\omega) = i_0, \ldots, X_{n-1}(\omega) = i_{n-1}\} \\ &= \mathcal{F}_n \cap \{\omega : \tau(\omega) = n\} \end{aligned}$$

by definition of $E_q^{(n)}(\omega)$. ∎

8.5.8 *Wald's equation.* Let τ be a stopping time with $E_q \tau < \infty$. Then $E_q \sum_{n=1}^{\tau} X_n = E_q \tau \cdot E_q X_1$.

Proof. We may write

$$\begin{aligned} E_q\left[\sup_n Y_n\right] &= \int_0^\infty P\left\{\sup_n Y_n > \lambda\right\} d\lambda \qquad \text{(see 6.2, Problem 7)} \\ &\le \sum_{n=1}^\infty P\left\{\sup_k Y_k > n\right\} \\ &= \sum_{n=1}^\infty \sum_{i_0, \ldots, i_{n-1}} P(A_{i_0} \cap \cdots \cap A_{i_{n-1}}) \\ &= \sum_{n=1}^\infty \sum_{i_0, \ldots, i_{n-1}} \sum_{k=1}^{n} P(E_q^{(k)} = n, \tau = n) \qquad \text{by 8.5.7} \end{aligned}$$

If the coordinate space is finite, then the theorem is an elementary consequence of the law of large numbers. Indeed, by (5.5.1),

$$-\frac{1}{n}\log p(X_0,\dots,X_{n-1}) = -\frac{1}{n}\sum_{i=0}^{n-1}\log p(X_i\mid X_0,\dots,X_{i-1}) = -\frac{1}{n}\sum_{i=0}^{n-1}\sum_{a\in A}\mathbf{1}_{\{X_i=a\}}\log p(a\mid X_0,\dots,X_{i-1}).$$

Since the $\{X_i=a\}$, $a=0,1,\dots$, are disjoint and the sum over i is finite,

$$-\frac{1}{n}\log p(X_0,\dots,X_{n-1}) = -\sum_{a\in A}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{1}_{\{X_i=a\}}\log p(a\mid X_0,\dots,X_{i-1}).$$

In the general case, assume without loss of generality that the numbers $p_i = P(A_i)$ are strictly decreasing (if necessary, we can replace Ω by the coordinate space). We then have $p_i \ge 1/i$ for all i, and if $q_i > 1/i$, then $q_1 + \cdots + q_i$ is strictly greater than 1. Hence, if $\kappa(i)$ is a function on the positive integers, then the nonnegative integers $i - \kappa(i)$ are nonnegative except, possibly, for a finite set of i. Thus by (5.5.2),

$$\sum_{i=1}^{\infty}\frac{1}{i}\left[\sum_{a\in A}\mathbf{1}_{\{X_{i-\kappa(i)}=a\}}\log p(a\mid X_{i-\kappa(i)},\dots,X_{i-1})\right] = \sum_{a\in A}\sum_{i:\,i-\kappa(i)\ge 0}\frac{1}{i}\mathbf{1}_{\{X_{i-\kappa(i)}=a\}}\log p(a\mid X_{i-\kappa(i)},\dots,X_{i-1})$$

by disjointness of the $\{X_{i-\kappa(i)}=a\}$, $a=0,1,\dots$.

Also by disjointness,

$$\sum_{n=1}^{\infty}\left|\sum_{i:\,i-\kappa(i)=n}\mathbf{1}_{\{X_{i-\kappa(i)}=a\}}\right| \le \sum_{i:\,i-\kappa(i)\ge 0}\mathbf{1}_{\{X_{i-\kappa(i)}=a\}} \le \sum_{i=0}^{\infty}\mathbf{1}_{\{X_i=a\}} = N_a.$$

Therefore,

$$E\left[\sup_n Y_n\right] \le \sum_{i=1}^{\infty}\frac{1}{i}\log\frac{1}{q_i} + \sum_{a\in A}\sum_{n=1}^{\infty}\frac{1}{n}\sum_{j:\,j\le N_a}p_j.$$

The second series is $\sum_{a\in A}\sum_{j=1}^{\infty}\frac{1}{j}\#\{n : n\ge j,\ n\le N_a\}$, where $\#(S)$ is the number of nonnegative integers in the set S. If $N_a = 0$ or $N_a = j-1$, then the last series converges, and $\#(\cdot) \le j$. If $N_a \ge j$, then $\#(\cdot) = N_a - j + 1$.

Choose a positive integer K such that $\sum_{j=K}^{\infty}\frac{1}{j}\log\frac{1}{p_j} < \frac{1}{2}$. For all $a \in A$ with $N_a \ge K$,

$$\sum_{j=K}^{N_a}\frac{1}{j} \le \log N_a - \log(K-1) \le \log N_a - \log\frac{K}{2},$$

and

$$\sum_{j=K}^{N_a}\frac{1}{j} \ge \log(N_a+1) - \log K \quad \text{implies} \quad \log N_a \ge \log\frac{K}{2}.$$

Therefore, $f_i(x) \le 2\log e \cdot g_i$, $i = 1, 2, \ldots$, hence

$$\sum_{i=1}^{2^n} f_i(x)\mu_i \le 2\sum_{i=1}^{2^n} \mu_i \log e \cdot g \sum_{i=1}^{2^n} \mu_i$$
$$+ \sum_{i=1}^{\infty} p_i \log \frac{1}{p_i} = H \quad \text{since} \quad \sum_{i=1}^{\infty} p_i = 1$$
$$= 2H(\mathcal{A}) < \infty.$$

We may now prove the main result:

8.5.9 Shannon–McMillan Theorem. Let T be a two-sided shift transformation with ergodic measure-preserving random variable X_0. Assume that $H(\mathcal{A}) < \infty$, where $\mathcal{A}$ is the atomic partition of the sequence $\{X_n\}$.

If $Z_n = -\frac{1}{n} \log p(X_0, \ldots, X_{n-1})$, there is a constant random variable Z such that $E|Z_n - Z| \to 0$ and $Z_n \to Z$ almost everywhere as $n \to \infty$. In particular, if H is ergodic, $Z = H$ a.e.

Proof. If the random variables Y_n are defined as in 8.5.4, we have

$$Z_n = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i \mid X_0, \ldots, X_{i-1})$$
$$= \frac{1}{n} \sum_{i=1}^{n-1} h_{\mathcal{A}}(T^i x)$$
$$= \frac{1}{n} \sum_{i=1}^{n-1} g(T^i x) - \frac{1}{n} \sum_{i=0}^{n-1} h_{\mathcal{A}}(T^i x) - g(T^i x)]. \tag{1}$$

By 8.5.6, 8.5.7, and 8.5.8, the first series in (1) converges to an integrable invariant function Z a.e. and in L^1. By 8.5.9, $E(Z) = E(g) = H$, and by 8.5.6, $\lim_{n\to\infty} E(|Z_n - Z|) = 0$. The second series converges to 0 in L^1 since $\|g \circ T^i - g\|_1 \to 0$ by 8.5.6, and $|E(Z^{(n)}) - E(Z)| = \|Z_n - Z\|_1$ by 7.6.12.

Thus it remains to show $Z = g$ a.e., all that remains is to show that $Z_n \to Z$ a.e., and this will follow if we can show that the second series in (1) converges to 0 a.e. Use the very precise integer N, the series Σ bounded in absolute value by

$$\frac{1}{n} \sum_{i=0}^{N} |g(T^i x) - g(T^i x)| + \frac{1}{n} \sum_{i=N}^{n-1} g(T^i x) - g(T^i x)|. \tag{2}$$

The first series approaches 0 as $n \to \infty$ since $\Pi(\mathcal{A}^n \mid \mathcal{P}^n) \le \Pi(\mathcal{A} \mid \mathcal{P})$ is integrable. Hence for each $\varepsilon > 0$ we can find n so that

$$G_n = \sup\{f(\cdot, \cdot, \ldots) : \ldots \in A\} < \varepsilon,$$

where $A = \operatorname{supp}(\eta_n)$, which is measurable by (3.3.8). Also since $\eta_n \to \delta_{(x,x,\ldots)}$ vaguely, we have $G_n < 0$ a.e. as $n \to \infty$, hence $\int G_n \, d\mu < 0$ by the dominated convergence theorem. But the second series in (3.2) (call it A_n) is bounded by $n^{-1} \sum_{k=1}^{n} \int G_k \, d\mu$, which converges to $\int G_n \, d\mu$ as well, and so in L^1, by Kolmogorov and (3.3.7).

Finally, given $\varepsilon > 0$, $\delta > 0$, we have $P(|Y_n - Y| \ge \varepsilon) < \varepsilon \cdot \delta / 2$ by Chebyshev's inequality, and by (3.3.9), $P(|Y_n - Y_m| \ge \varepsilon) < \delta/2$ if we choose N such that $P(|Y_n - Y| \ge \varepsilon) < \delta/2$ for $n \ge N$. Since ε and δ are arbitrary, it follows that $Y_n \to Y$ a.e. $\square$

Since the Shannon-McMillan theorem involves only the random variables Y_n, $n > 0$, it holds equally well for a measure-preserving shift. For an intuitive sense of how to prove this formally, see the discussion at the end of Problem 2 in §5.

Problems

1. (a) In the Shannon–McMillan theorem, give an example in which the limit random variable Y is not constant, but Y is not equal to η.

   (b) In the Shannon–McMillan theorem, give an example in which Y is not constant and (of course) Y cannot be equal to η.

2. If the coordinate space $\mathcal{A}$ is finite and $\mathcal{P} = \mathcal{B}$, show that the random variables $Z_n^{\mathcal{P}}$ are uniformly integrable, thus $Z_n \to Z$ in L^1.

3. [A short proof of a special case of the Shannon–McMillan theorem (Billingsley, [1965b]).] Let $X_1, X_2, \ldots$ be a stationary sequence of discrete random variables, with $H(X_1) < \infty$. In this problem, entropy will be expressed using natural logarithms (for convenience). Define an auxiliary process $Y_n = p(X_n, \ldots, X_1)$ by

$$p(X_n, \ldots, X_1) = p(X_1, \ldots, X_n) \, p(X_2, \ldots, X_{n+1} \mid X_1, \ldots, X_n) \cdots$$
$$p(X_{n-1}, \ldots, X_{2n-2} \mid X_{n-2}, \ldots, X_1) \quad (n = 1, 2, \ldots).$$

   Now note that $\log Y_n = \log Y_{n-1} + \log p(X_{2n-1} \mid X_{n-1}, \ldots, X_1) - \log p(X_{n-1} \mid X_{n-2}, \ldots, X_1)$, so that

$$E\left(\log \frac{Y_n}{Y_{n-1}} \,\middle|\, \mathcal{F}_{n-1}\right) = E\left(\log \frac{Y_n}{p_{n-1}} \,\middle|\, \mathcal{F}_{n-1}\right) = H\left(\log \frac{\eta_n}{p_{n-1}}\right)$$

   where $p_{n-1} = p(X_{n-1}, \ldots, X_1)$, $\eta_n = p(X_{2n-1}, \ldots, X_n)$.

(a) Show that $E\{a_{n+m}(x)\} \le 1$ and

$$E\left\{-\frac{1}{n}\ln\frac{a_{n+m}(x)}{p(x)}\right\} = H(X_0,\ldots,X_{m-1}) + \frac{1}{n}H(X_m \mid X_0,\ldots,X_{m-1}) - H(X_0,\ldots,X_{m-1}).$$

Thus

$$\limsup_{n\to\infty} E\left\{\frac{1}{n}\ln\frac{a_{n+m}(x)}{p(x)}\right\} \le \lim_{n\to\infty}\frac{1}{n}H(X_0,\ldots,X_{m-1}) + \lim_{n\to\infty}\frac{1}{n}H(X_m \mid X_0,\ldots,X_{m-1}) - H(X_0,\ldots,X_{m-1}) = 0.$$

(b) Assume $\{X_n\}$ ergodic. Recall that the associated shift transformation is ergodic. Show that as $n \to \infty$, $\frac{1}{n}\ln\frac{a_{n+m}(x)}{p(x)}$ converges in mean and in L^1 to $H(X_m \mid X_0,\ldots,X_{m-1})$.

(c) Let H be the entropy of the ergodic sequence $\{X_n\}$. Given $\varepsilon > 0$, choose an integer m such that

$$\left|\frac{H(X_0,\ldots,X_{m-1})}{m} - H\right| < \frac{\varepsilon}{2}$$

and

$$|H(X_m \mid X_0,\ldots,X_{m-1}) - H| < \frac{\varepsilon}{2}.$$

Show that $\delta\{-\frac{1}{n}\ln p(x_0,\ldots,x_{n-1}) - H\} < \varepsilon$ for sufficiently large n, thus establishing AEP convergence in the Shannon–McMillan theorem under the hypothesis of ergodicity.

8.6 Entropy of a Transformation

We define in this section the notion of entropy of a measure-preserving transformation, and show that two isomorphic transformations have the same entropy. For Bernoulli shifts this answers in the negative the Bernoulli shift problem: two Bernoulli shifts with the same entropy are isomorphic.

8.6.1 Definitions

(a) Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{A} = \{A_1,\ldots,A_k\}$ a finite subfield of $\mathcal{F}$. Each set A_i is a cylinder with base in $A_1,\ldots,A_k$. The entropy of $\mathcal{A}$ is by definition

$$H(\mathcal{A}) = -\sum_{i=1}^{k} P(A_i)\log P(A_i).$$

(b) If $\mathcal{A}$ is a finite field of subsets of Ω and $A_1, \ldots, A_n$ is a measurable partition of Ω (that is, $A_i \in \mathcal{A}$ for each i), then

$$H(\mathbb{P}; \mathcal{A}) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P(A_i \cap B_j) \log P(B_j \mid A_i).$$

There is a further striking way in which definitions (a) and (b) of the discrete random variables X and Y are defined by $X = i$ on A_i, $Y = j$ on B_j, we have $H(\mathbb{P}; \mathcal{A}) = H(X)$ and $H(\mathbb{P}; \mathcal{B}) = H(Y)$.

Since we consider only finite subfields of $\mathcal{F}$, $H(\mathbb{P}; \mathcal{A})$ and $H(\mathbb{P}; \mathcal{B} \mid \mathcal{A})$ are always finite and the properties of entropy given in 6.3.3 can be rewritten in terms of finite subfields.

6.4.2 Theorem Let $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ be finite subfields of $\mathcal{F}$. We have the following properties (where $\mathcal{A} \vee \mathcal{B}$ denotes the smallest (finite) field containing $\mathcal{A}$ and $\mathcal{B}$):

(a) $H(\mathbb{P}; \mathcal{A} \vee \mathcal{B}) \leq H(\mathbb{P}; \mathcal{A}) + H(\mathbb{P}; \mathcal{B})$ with equality iff $\mathcal{A}$ and $\mathcal{B}$ are independent.

(b) $H(\mathbb{P}; \mathcal{A} \vee \mathcal{B} \mid \mathcal{C}) = H(\mathbb{P}; \mathcal{A} \mid \mathcal{C}) + H(\mathbb{P}; \mathcal{B} \mid \mathcal{A} \vee \mathcal{C})$.

(c) Consequently, $H(\mathbb{P}; \mathcal{A}) \leq H(\mathbb{P}; \mathcal{B})$ if $\mathcal{A} \subseteq \mathcal{B}$.

(d) $H(\mathbb{P}; \mathcal{A} \mid \mathcal{B}) \leq H(\mathbb{P}; \mathcal{A})$, with equality iff $\mathcal{A}$ and $\mathcal{B}$ are independent.

(e) If $\mathcal{A}$ has n atoms, $H(\mathbb{P}; \mathcal{A}) \leq \log n$, with equality iff each atom of $\mathcal{A}$ has probability $1/n$.

(f) $H(\mathbb{P}; \mathcal{A} \vee \mathcal{C} \mid \mathcal{B} \vee \mathcal{C}) \leq H(\mathbb{P}; \mathcal{A} \mid \mathcal{B})$, with equality iff $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{C}$.

(g) $H(\mathbb{P}; \mathcal{A} \vee \mathcal{B} \mid \mathcal{C}) = H(\mathbb{P}; \mathcal{A} \mid \mathcal{C}) + H(\mathbb{P}; \mathcal{B} \mid \mathcal{C}) - I(\mathbb{P}; \mathcal{A} \wedge \mathcal{B}) + H(\mathbb{P}; \mathcal{A} \mid \mathcal{B} \wedge \mathcal{C}) + H(\mathbb{P}; \mathcal{B} \mid \mathcal{A} \wedge \mathcal{C})$.

(h) Consequently, $H(\mathbb{P}; \mathcal{A} \vee \mathcal{B}) \geq H(\mathbb{P}; \mathcal{A})$ if $\mathcal{B} \supseteq \mathcal{A}$.

(i) $H(\mathbb{P}; \mathcal{A} \vee \mathcal{B}) \leq H(\mathbb{P}; \mathcal{A}) + H(\mathbb{P}; \mathcal{B})$, with equality iff $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{A} \wedge \mathcal{B}$.

(j) Consequently, $H(\mathbb{P}; \mathcal{A}) \leq H(\mathbb{P}; \mathcal{B})$ if $\mathcal{A} \subseteq \mathcal{B}$.

Entropy is often said to be a measure of uncertainty. We give a few examples to justify this statement.

1 Consider two finite subfields, $\mathcal{A}$ and $\mathcal{B}$, of $\mathcal{F}$, and assume that $\mathcal{B} \supseteq \mathcal{A}$. Then there is each atom of $\mathcal{A}$ is a union of atoms of $\mathcal{B}$. If we know that the outcome ω is in the atom A_i of $\mathcal{A}$, we still do not know which atom of $\mathcal{B}$ contained ω, but knowing that it is in the atom B_j of $\mathcal{B}$ is enough

to tell us which atom of $\mathcal{A}$ contains ω. This means that individually the uncertainty is greater for the subfield $\mathcal{A}$ than for the subfield $\mathcal{B}$. This agrees with the inequality $H(\mathcal{A}) \geq H(\mathcal{B})$.

2. If we tell a Laplacian uncertainty about the outcome is intuitively greatest when the die is unbiased, and is when the known correlation between the X_n is zero, $\rho = 0$. This is exactly what $H(\mathcal{A}) = \log 6$ states.

3. The inequality $h(\mathcal{A}) \leq H(\mathcal{A})$ states that the more we know the less the uncertainty.

8.6.3 Remarks. Entropy is invariant under measure preserving isomorphisms. That is, if T is a measure-preserving transformation on $(\Omega, \mathcal{F}, P)$, and if $\mathcal{A}$ and $\mathcal{B}$ are partitions which are finite or countable, and if $T^{-1}\mathcal{A}$ and $T^{-1}\mathcal{B}$ are refinements of $\mathcal{A}$, then $H(\mathcal{A} \mid T^{-1}\mathcal{A}) = H(\mathcal{B} \mid T^{-1}\mathcal{B})$. The measure preserving property of T implies that $H(T^{-1}\mathcal{A}) = H(\mathcal{A})$ and $H(T^{-1}\mathcal{A} \mid T^{-n}\mathcal{A}) = H(\mathcal{A} \mid \bigvee_{k=1}^{n} T^{-k}\mathcal{A})$, $n = 1, 2, \ldots$.

Because of the correspondence between measure-preserving transformations and stationary sequences, the entropy $H(X_0 \mid \mathcal{P})$ of a stationary sequence can also be described in terms of measure preserving transformations.

8.6.4 Definition. If $A_1, \ldots, A_k$ are the atoms of $\mathcal{A}$, and X is the random variable defined by $X = i$ on A_i, the sequence $X_1 = X, X_2 = X \circ T, \ldots, X_n = X \circ T^{n-1}, \ldots$ is a stationary sequence and

$$H\left(T^{-1}\mathcal{A} \mid \bigvee_{n=1}^{\infty} T^{-n}\mathcal{A}\right) = H(X_0 \mid X_1, X_2, \ldots, X_{-n}, \ldots)$$

By 8.6.3, $\lim_{n\to\infty} H\left(T^{-1}\mathcal{A} \mid \bigvee_{k=1}^{n} T^{-k}\mathcal{A}\right) = \lim_{n\to\infty} H(X_0 \mid X_1, X_2, \ldots, X_{-n}) = H(X_0 \mid X_1, X_2, \ldots)$.

The entropy of a measure-preserving transformation T with respect to a given partition $\mathcal{A}$ is defined as

$$H(\mathcal{A}, T) = \lim_{n\to\infty} H\left(\mathcal{A} \mid \bigvee_{k=1}^{n} T^{-k}\mathcal{A}\right)$$

The quantity $U(X_n) = \lim_{n\to\infty} \frac{1}{n} U(X_1, \ldots, X_n)$ proved in 8.6.1 can be written in terms of $H(\mathcal{A}, T)$.

8.6.5 Theorem.

(a) $H(\mathcal{A}, T) = \lim_{n\to\infty} H\left(\mathcal{A} \mid \bigvee_{k=1}^{n} T^{-k}\mathcal{A}\right)$.

(b) $H(\mathcal{A}, T) = \lim_{n\to\infty} H\left(\mathcal{A} \mid \bigvee_{k=1}^{\infty} T^{-k}\mathcal{A}\right)$.

Proof. (a) is a consequence of 5.3.2.

(b) By 5.3.2(f) we have $E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) \geq E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n}\right) \geq 0$ and thus the limit $\lim_{n\to-\infty} E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n}\right)$ exists. We set

$$\begin{aligned} \alpha\left(\bigvee_{-\infty}^{-1}\mathcal{A}\right) &= \alpha\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{-1}\right) = \alpha\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) && \text{by 5.6.2(b)} \\ &= E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) - E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{-1}\right) && \text{by 5.6.1} \\ &= E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) + E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n}\right) - \dots \\ &\qquad E(\mathcal{A}\,|\,\mathcal{I}^{-1}\mathcal{A}) + E(\mathcal{I}^{-1}\mathcal{A}) \end{aligned}$$

The argument in 5.6.3 shows that

$$\lim_{n\to-\infty} E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) = \lim_{n\to-\infty} E\left(\mathcal{I}^{-1}\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) = E(\mathcal{A}\,|\,\mathcal{I}^{-1}\mathcal{A}).$$

6.6.6 *Corollary.* If $\mathcal{I}$ is an invertible countably, one-to-one, ergodic, and $\mathcal{I}^{-1}$ absolutely continuous transformation, then

$$E(\mathcal{A}\,|\,\mathcal{I}^{-1}\mathcal{A}) = \lim_{n\to-\infty} E\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) = \lim_{n\to-\infty} E\left(\mathcal{I}^{n}\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right).$$

Notice that we have written this corollary in terms of ascending sequences. Since $\mathcal{I}$ is invertible, we can define the two-sided stationary sequence $X_n = X_0\mathcal{I}^n$, $n = 0, \pm 1, \pm 2, \dots$, where $X = X_0$ is the identity. The corollary then asserts that $E(X_0 \circ \mathcal{I}) = \lim_{n\to-\infty} E(X_0 \circ \mathcal{I} \mid X_{-1}, \dots, X_n) = \lim_{n\to-\infty} E(X_1 \mid X_0, \dots, X_n)$. (Here we must recall that in the definition of $E(\cdot\,|\,\mathcal{X}_0^{-1})$ we count on the time backwards.)

Proof. By 5.6.5,

$$\begin{aligned} \alpha\left(\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) &= \alpha\left(\mathcal{I}^{n-1}\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) \\ &= E\left(\mathcal{I}^{n-1}\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right) - E\left(\mathcal{I}^{n}\mathcal{A}\,\middle|\,\bigvee_{-\infty}^{n-1}\right). \end{aligned}$$

and

$$H\left(\mathcal{V}_1^{n}\middle|\mathcal{V}_1^{n}\right) = H\left(\mathcal{V}_1^{n}\right) - H\left(\mathcal{V}_1^{n-1}\right) = \cdots = \sum_{k=1}^{n} H\left(\mathcal{V}_k\middle|\mathcal{V}_1^{k-1}\right).$$

A.6.7 *Remark.* If $\mathcal{A}$ and $\mathcal{B}$ are two finite subfields of $\mathcal{F}$, we have

(a) $H(\mathcal{A}\vee\mathcal{B},T) \le H(\mathcal{A},T) + H(\mathcal{B},T)$ if $\mathcal{A}\subset\mathcal{F}$, $\mathcal{B}\subset\mathcal{F}$.

(b) $H(\mathcal{A}\vee\mathcal{B},T) = H(\mathcal{A},T) + H(\mathcal{B},T)$ if $\mathcal{A}\subset\mathcal{F}$, $\mathcal{B}\subset\mathcal{F}$, and $\mathcal{A}$ and $\mathcal{B}$ are independent.

(c) $H(\mathcal{A},T) \le H(\mathcal{B},T) + H(\mathcal{A}|\mathcal{B})$.

Item (a) is an easy consequence of (b) and (c).

For (b), notice that

$$H\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}\vee\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}\right) = H\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}\right) + H\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}\right)$$

$$= nH\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}\right) + nH\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}\right).$$

In (c),

$$H\left(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{A}\middle|\bigvee_{k=0}^{n-1} T^{-k}\mathcal{B}\right)$$

$$\le \sum_{k=0}^{n-1} H\left(T^{-k}\mathcal{A}\middle|\bigvee_{j=0}^{n-1} T^{-j}\mathcal{B}\right)$$

$$= \sum_{k=0}^{n-1} H\left(T^{-k}\mathcal{A}\middle|\bigvee_{j=0}^{k} T^{-j}\mathcal{B}\right) = nH(\mathcal{A},T|\mathcal{B}).$$

We are now ready to define the entropy of a measure preserving transformation in such a way that it is invariant under conjugation. We first specify what we mean by conjugate measure preserving transformations.

8.6.8 *Definition (Isomorphic transformations).* Let $(\Omega_1,\mathcal{F}_1,P_1,T_1)$ and $(\Omega_2,\mathcal{F}_2,P_2,T_2)$ be two probability spaces with measure preserving transformations. $(\Omega_1,\mathcal{F}_1,P_1,T_1)$ and $(\Omega_2,\mathcal{F}_2,P_2,T_2)$ are isomorphic iff there exist $\Omega_1' = T_1$ and $\Omega_2' \in \mathcal{F}_2$, and a mapping $\Phi: \Omega_1' \to \Omega_2'$ such that

(a) $P(\Omega_1^*) = P(\Omega_2^*) = 1$.

(b) The mapping Ψ is one to one and onto; that is, if $B_1 \in \mathcal{B}(\Omega_1^*)$, $B_2 = \psi(B_1)$ $\in \mathcal{B}^*$, then $B_2 \in \mathcal{B}_2^*$.

(c) For any subset $B_1 \in \mathcal{B}_1$ of Ω_1^*, $P(\psi(B_1)) = P_1(B_1)$.

(d) $T_1(\Omega_1^*) \subset \Omega_1^*$ and $T_2(\Omega_2^*) \subset \Omega_2^*$.

(e) $\Psi T_1 \omega = T_2 \Psi(\omega)$ for each $\omega \in \Omega^*$.

Properties (b) and (c) state that the probability structures $(\Omega_1^*, \mathcal{B}_1^*, P_1)$ and $(\Omega_2^*, \mathcal{B}_2^*, P_2)$ are isomorphic [recall $\mathcal{B}_1^*$ (resp. $\mathcal{B}_2^*$) denotes the σ-field consisting of the subsets of Ω_1^* (resp. Ω_2^*) which are in $\mathcal{B}^*$ (resp. in $\mathcal{B}^*$)]. Property (a) is purely technical; it allows us to limit ourselves to what happens on Ω_1^* and Ω_2^*. Property (e) states that T_1 and T_2 are identical as transformations on Ω_1^* and Ω_2^* are compatible with the isomorphism Ψ. We could have defined the isomorphic structures $(\Omega_1^*, \mathcal{B}_1^*, P_1, T_1)$ and $(\Omega_2^*, \mathcal{B}_2^*, P_2, T_2)$ by imposing that $\Omega_1^* = \Omega_1$ and $\Omega_2^* = \Omega_2$, but nothing is changed in the definition of a probability space if we set ourselves on a set of measure zero. It is therefore natural to add the property (a) to the definition.

6.6.9 *Definition.* Let T be a measure-preserving transformation on a probability space $(\Omega, \mathcal{F}, P)$. The entropy of T is

$$h(T) = \sup_{\mathcal{A}} h(\mathcal{A}, T) \qquad \text{where the sup is taken over}$$

$$\mathcal{A} \text{ finite subfields } \mathcal{A} \subset \mathcal{F}.$$

6.6.10 *Theorem.* Let T_1 and T_2 be measure preserving transformations on $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ respectively. Suppose that $(\Omega_1, \mathcal{F}_1, P_1, T_1)$ and $(\Omega_2, \mathcal{F}_2, P_2, T_2)$ are isomorphic. Then $h(T_1) = h(T_2)$.

Proof. This result is a consequence of the following remarks (see also the notations of 6.6.8):

(a) The entropies of T_1 in Ω_1 and Ω_1^* are the same because $P(\Omega_1 - \Omega_1^*) = 0$. The same is true for the entropies of T_2 in Ω_2 and Ω_2^*.

(b) For any finite subfield $\mathcal{A} = \{A_1, \ldots, A_n\}$ of $\mathcal{F}_1^*$, $\Psi \mathcal{A} = \{\psi(A_1), \ldots, \psi(A_n)\}$ is a finite subfield of $\mathcal{F}_2^*$, and any finite subfield of $\mathcal{F}_2^*$ is of this form.

(c) If $\mathcal{A} = \{A_1, \ldots, A_n\}$ is a finite subfield of $\mathcal{F}_1^*$,

$$H(\Psi \mathcal{A}, T_2) = H(\mathcal{A}, T_1).$$

To conclude this section we give two results, which in certain cases allow us to compute $h(T)$ easily.

8.6.11 *Theorem.* If T is a measure preserving transformation on a probability space $\Omega, \mathcal{F}, P$ and $\mathcal{I}$ is the invariant σ-field of $\mathcal{F}$ (recall that the events $A \in \mathcal{F}$ with $T^{-1}A = A$ are invariant), then $P(A \cap T^{-n}A) \to P(A)^2$ iff T is ergodic. The independence of the $\{T^{-n}A : n = 0, 1, \ldots\}$ for $A = \{X_0 \in B\}$ is equivalent to the independence of the $X_n \circ \theta$ for $n = 0, 1, 2, \ldots$ by the measure preserving property of T.

Proof We have

$$\begin{aligned} P(A \cap T^{-n}A) &= \lim_{N \to \infty} \frac{1}{N} \sum_{m=0}^{N-1} P\left(\bigcap_{k=0}^{m} T^{-k+n}A\right) \qquad \text{by 8.6.5} \\ &\lim_{N \to \infty} \frac{1}{N} \left[P(A) + P(T^{-n}A \cap A) + \cdots + P(T^{-N+1}A \cap \cdots \cap A)\right] \\ &\qquad \text{as } N \to \infty \text{ this} \\ &= \lim_{N \to \infty} \frac{1}{N} \left[P(A) + P(A \cap T^{-n}A) + \cdots + P(A \cap T^{-N+1}A)\right] \qquad \text{by 8.6.1} \\ &= P(A) \cdot P(A) \end{aligned}$$

We will now show that if $\bigcap_{n=0}^{\infty} T^{-n}A$ generates $\mathcal{F}$, we have $P(T^{-n}A \cap B) \to P(A)P(B)$. We claim by proving two lemmas.

8.6.12 *Lemma.* Let T be an invertible measure preserving transformation on the probability space $\Omega, \mathcal{F}, P$ and $\mathcal{A} \subset \mathcal{F}$ be a finite subfield of $\mathcal{F}$. We have, for every $n = 0, 1, \ldots$

$$P\left(\bigcap_{m=-n}^{n} T^{-m}A_m\right) = P\left(\bigcap_{m=-n}^{n} T^{-m}A_{m+n}\right)$$

Proof.

$$\begin{aligned} P\left(\bigcap_{m=-n}^{n} T^{-m}A_m\right) &= \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P\left(\bigcap_{m=-n}^{n} T^{-m-k}A_m\right) \qquad \text{by 8.6.5 applied} \\ &= \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P\left(\bigcap_{m=-n}^{n} T^{-m}A_{m+k}\right) \qquad \text{by 8.6.1} \\ &= \lim_{N \to \infty} \frac{1}{N} \sum_{k=n}^{N+n-1} P\left(\bigcap_{m=-n}^{n} T^{-m}A_{m+k}\right) \\ &= P\left(\bigcap_{m=-n}^{n} T^{-m}A_{m+n}\right) \end{aligned}$$

8.5.11 *Lemma.* Let $\mathcal{A}$ and $\mathcal{B}$ be two finite subfields of $\mathcal{F}$. Assume that both $\mathcal{A}$ and $\mathcal{B}$ have N atoms, designated by $A_1, \ldots, A_N$ and $B_1, \ldots, B_N$, respectively. Then for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $\sum_{i=1}^{N} P(A_i \triangle B_i) < \delta$ implies $|H(\mathcal{A}, \mathcal{B}) - H(\mathcal{A}, P)| \leq \varepsilon$.

Proof. Let $C_i = \bigcap_{j=1}^{N} (A_i \cap B_j)$ and $C_{N+1} = \Omega - \bigcup_{i=1}^{N} C_i$. We have $P(C_{N+1}) < \delta$. If $P(C_i) = 0$, then $P(A_i \triangle B_i) = 0$ for all i. In this case $H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A} \mid \mathcal{C})$. Hereafter we can assume that $P(C_i) > 0$. If we apply (8.5.4) to the probability $P(\cdot \mid C_i) = P(\cdot \cap C_i)/P(C_i)$ on $\Omega = C_i$, we see that

$$\sum_{i=1}^{N} P(C_i) \log P(C_i) \leq \log 2.$$

Replacing $P(C_i)$ by $P(C_i, C_{N+1}) + P(C_i)$ we get

$$-\sum_{i=1}^{N} P(C_i) \log P(C_i) < (1 - P(C_{N+1})) \log 2 - (1 - P(C_{N+1})) \log (1 - P(C_{N+1}))$$
$$\leq 2 \log 2 - \delta \log \delta.$$

If $\delta < 1/2$, and therefore P is on the finite subfield generated by the C_i, $i = 0, \ldots, N$, we have

$$H(\mathcal{A} \mid \mathcal{B}) = -\sum_{i,j=1}^{N} P(A_i \cap B_j) \log P(A_i \mid B_j) + \sum_{j=1}^{N} P(B_j) \log P(B_j) - \sum_{i=1}^{N} P(A_i) \log P(A_i)$$
$$\leq -\sum_{i,j=1}^{N} P(A_i \cap B_j) \log P(A_i \mid B_j) + \sum_{j=1}^{N} P(B_j) \log P(B_j)$$
$$\leq -\sum_{i,j=1}^{N} P(C_i \cap B_j) \log P(C_i \mid B_j) + \sum_{j=1}^{N} P(B_j) \log P(B_j) + 2 \log 2 - \delta \log \delta$$
$$= H(\mathcal{C} \mid \mathcal{B}) + 2 \log 2 - \delta \log \delta.$$

If $\delta < 1/2$, we can choose δ in such a way that $H(\mathcal{C} \mid \mathcal{B}) < \varepsilon$. We then have by (8.5.10)

$$H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A} \mid \mathcal{C}) + H(\mathcal{C} \mid \mathcal{B}) < H(\mathcal{A} \mid \mathcal{C}) + \varepsilon.$$

Since $\mathcal{A} \subset \sigma(\mathcal{C}, \mathcal{N})$, where $\mathcal{N} = \mathcal{A} \cap \mathcal{N}(P)$, by (8.5.7)(g) we have $H(\mathcal{A} \mid \mathcal{C}) \leq H(\mathcal{A} \mid \mathcal{B}, \mathcal{N}) = 0$. Therefore $H(\mathcal{A}, P) < H(\mathcal{A} \mid \mathcal{B}) + \varepsilon < 2\varepsilon$ for sufficiently small δ. We can similarly prove that $H(\mathcal{B} \mid \mathcal{A}) < \varepsilon$ for sufficiently small δ. ∎

8.5.12 *Theorem* (Kolmogorov–Sinai). If T is an invertible measure-preserving transformation on the probability space $(\Omega, \mathcal{F}, P)$, and if the finite subfield $\mathcal{A}$ is such that the σ-field generated by the infinite $\bigcup_{n} T^n \mathcal{A}$ is equal to $\mathcal{F}$, then $h(T) = H(\mathcal{A} \mid T_+)$.

Proof. Let $\mathcal{A}$ be a finite subfield of $\mathcal{F}$ with atoms $A_1, \ldots, A_N$. Given any $\varepsilon > 0$, we can find $C_1, \ldots, C_M \in \bigcup_{n} T^n \mathcal{A}$ such that $P(A_i \triangle C_i) < \delta$

$I = 1, \ldots, r$. Let N be such that all the C_i are in $\bigvee_{i=0}^{N-1} T^{-i}\mathcal{P}$. Then $C_i = T^{-1}C_{i-1}$, $C_i = T^{-i}C_0$, and denote by $\mathcal{C}$ the subfield of $\mathcal{F}$ having the C_i as atoms. We have $\sum_{i=1}^{r} |\nu(C_i) - \mu(C_i)| < \delta/2$. Given $\varepsilon > 0$, Rohlin allows us to choose $\delta > 0$ such that $H(\mathcal{C}|\mathcal{D}) < \varepsilon/2$ implies $H(\mathcal{D}|\mathcal{C}) < \varepsilon$. And we have

$$\begin{aligned} H(\mathcal{P}, T\mathcal{P}, \ldots, T^{N-1}\mathcal{P}) &< H(\mathcal{C}, T\mathcal{P}, \ldots, T^{N-1}\mathcal{P}) + \varepsilon \\ &< H\left(\bigvee_{i=0}^{N-1} T^{-i}\mathcal{P}\right) + \varepsilon \\ &= H(\mathcal{P}, T\mathcal{P}, \ldots, T^{N-1}\mathcal{P}) + \varepsilon \quad \text{by 8.6.12} \end{aligned}$$

Since this is true for any finite subfield $\mathcal{P}$ of $\mathcal{F}$ we have $h(T) = h(T, \mathcal{C}) < h(T, \mathcal{P}) + \varepsilon$. □

8.6.15 *Corollary.* If T is an invertible measure preserving transformation on a probability space $(\Omega, \mathcal{F}, P)$, and $\mathcal{P}$ is a finite subfield of $\mathcal{F}$ such that the sub-σ-fields $\mathcal{P}_n = \bigvee_{i=-n}^{n} T^{-i}\mathcal{P}$ are independent and the subfield they generate is equal to $\mathcal{F}$, we have $h(T) = h(T, \mathcal{P})$.

8.7 Bernoulli Shifts

We can now apply the results of 8.6 to compute the entropy of a Bernoulli shift.

8.7.1 *Definition (Bernoulli shifts).* Let S be the set $\{1, \ldots, k\}$, let $\Omega = S^{\mathbb{Z}}$ be the σ-field of all the subsets of Ω, and $P = (p_1, \ldots, p_k)^{\mathbb{Z}}$ be the probability measure on $(\Omega, \mathcal{F})$ such that $P\{\omega : \omega(i) = s\} = p_s$, $i = 0, \pm 1, \ldots$, $s = 1, \ldots, k$. The quadruple $(\Omega, \mathcal{F}, P, T)$ consisting of all doubly infinite sequences $\omega = (\ldots, \omega(-1), \omega(0), \omega(1), \ldots)$ of symbols $\omega(i) = s$ with probability p_s, the σ-field $\mathcal{F}$ generated by the cylinders $C = \{\omega : \omega(-n) = s_{-n}, \ldots, \omega(n) = s_n\}$, where the s_i can be any elements of S, the probability of such a cylinder is $P(C) = \prod_{i=-n}^{n} p_{s_i}$, and the shift transformation T defined by $T\omega(i) = \omega(i+1)$ (the $\omega(i)$ are the coordinate random variables $\omega(i) = \omega_i$ generally ω_i are not independent) is called a *Bernoulli shift*.

The Bernoulli shift with distribution π is the transformation T_π defined by

$$\omega_\pi(T_\pi\omega)(i) = \omega_\pi(\omega)(i+1)$$

(The transformation T_π shifts the coordinates of ω to the left.) Here, T_π is measure-preserving and invertible because if C is a cylinder, $T^{-1}C$ and TC are also cylinders and have the same probability as C.

8.7.2 *Theorem (Entropy of a Bernoulli shift).* The Bernoulli shift T_π with distribution $\pi = (p_1, \ldots, p_k)$ has entropy $H(T_\pi) = -\sum_{i=1}^{k} p_i \log p_i$.

Proof. Let $\mathcal{A}$ be the finite subfield generated by the partition η. Thus $\mathcal{A}$ is generated by the cylinders $[x_0, \ldots, x_{n-1}]$, $x_i = 0, \ldots, k-1$. The subfield $T^{-n}\mathcal{A}$ is generated by the partitions α_n. Therefore the subfields $T^{-n}\mathcal{A}$, $n = 0, 1, \ldots$, are independent and generate $\mathcal{F}_\eta$, and according to 6.5.13

$$H(\alpha_n) = H(\alpha) = -\sum_{i=1}^{k} p_i \log p_i .$$

Example. The random variables $\xi_n : \Omega \to \{1, \ldots, k\}$ of the distribution p given on Ω are one of the symbols, the probability space $(\Omega_\eta, \mathcal{F}_\eta, P_\eta)$ are all isomorphic to the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$, where λ is Lebesgue measure and the shift automorphism T. $\square$

We recall that two probability spaces $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ are isomorphic if (properties a), b), for all $A \in \mathcal{F}_1$ are satisfied

6.2.1 Theorem. For any integer $k \geq 2$ and any distribution $p = (p_0, \ldots, p_{k-1})$, $p_i > 0$, $\sum p_i = 1$, the probability spaces $(\Omega_\eta, \mathcal{F}_\eta, P_\eta)$ is isomorphic to the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$, where λ is Lebesgue measure.

If $p_i = \frac{1}{k}$ for all i, the proof of the isomorphism of the probability spaces $(\Omega_\eta, \mathcal{F}_\eta, P_\eta)$ is essentially contained in the exercises and cannot be considered to the end of the section.

Exercise. If the distribution p gives the same probability $\frac{1}{k}$ to each of the k symbols, the proof is easy. For each point x of the unit square we write the expansion in base k of first and y and finite sums of x

$$x = .x_0 x_2 x_4 \ldots, \qquad y = .x_1 x_3 x_5 \ldots$$

and we define an $\ldots$

$$\varphi(x, y) = \ldots$$

If one of the coordinates of x is of the form a/k^n, it has two expansions in base k since $1/k^n = \sum_{i=n+1}^{\infty} (k-1)/k^i$, and φ is not well defined. We have to first remove such points from the unit square, but we are still left with a set E of measure 0. It is easy to check that $\varphi(E)$ has measure 0, and φ is an isomorphism from E to φE.

If the p_i are not all equal, we have to modify the proof to make sure that $P_\eta(C) = \mu(\varphi(C))$ for all Borel subsets C of E. We construct the image $\ldots, x_{-1}, x_0, x_1, \ldots$ of a point x of $[0,1]^{\mathbb{Z}}$ in the following way for

on the square $[0,1]^2$ in a vertical open band $A_{j,1}, \ldots, A_{j,n_j}$ with width proportional to $p_{j,1}, p_{j,2}, \ldots, p_{j,n_j}$ so that

$$A_{j,i} = [\text{vertical lines at } x = a_{j,i}], \qquad A_{j,i} = \left\{ (x,y) : a_{j,i} \sum_{m=1}^{i-1} p_{j,m} < x < a_{j,i} \sum_{m=1}^{i} p_{j,m} \right\}$$

where the $p_{j,i}$ are all equal to $1/2^j$, finding the first member of the sequence $x_1, x_2, \ldots$ in base 2 requires crossing $[0,1]^2$ in several vertical bands. If the point x is in $A_{j,i}$, we suppose $\omega_j = i$ for the jth coordinate of ω. If the point x is in the vertical boundaries of more than one $A_{j,i}$, then we do not define ω_j.

We now divide the band $A_{j,i}$ into n_j vertical open bands of width proportional to $p_{j,1}, p_{j,2}, \ldots, p_{j,n_j}$ and if the point x is in the ith band we suppose $\omega_{j+1} = i$ for the first coordinate of ω beyond the n_j we define ω_j when the point x falls on the vertical boundaries of one of these new bands. We denote this procedure by defining the values of ω_{j+1} for $x = (0,1)$. We define the ω_{j+1} for $x = (1,2)$ we use the same procedure but we divide the square $[0,1]^2$ into horizontal open bands.

Note that ω_j is not defined for the points x that form the boundary of one of the bands so defined. Since there are only a countable number of these bands, the function ϕ is defined on a set E of Lebesgue measure 1. The function ϕ is not one-to-one. The sequences ω not containing an infinite number of consecutive 1's or an infinite number of consecutive 0's are never images of elements of E. Let A_0 be the set $[\omega_j = 0]$ for all $j \geq n$, and B_0 be the set $[\omega_j = 1]$ for all $j \geq n$, and define the sets A_i and B_i similarly by replacing the condition $\omega_j = 0$ by $\omega_j = i$. The set of elements of Ω_0 that are not images under ϕ of elements of E is $\bigcup_{i=0}^{\infty} \phi(A_i \cup B_i)$. Each A_i and each B_i is in $\mathcal{F}_n$ and has measure 0, therefore $\Sigma' = \phi E$ is measurable, also of P_0-measure one. Thus the restriction of ϕ to Σ is one-to-one.

By construction, the image of an open rectangle contained in Ω' [illegible] $(a_1, b_1) \times \cdots \times (a_n, b_n) \times [0,1] \times [0,1] \times \cdots$ and has P_0-measure equal to Lebesgue measure of the rectangle. Since the rectangles generate the Borel σ-field on the unit square, and the cylinders generate $\mathcal{F}_0$, the probability space $(\Omega, \mathcal{F}_0, P_0)$ is isomorphic to the unit square with Lebesgue measure.

Now that we know that essentially all the probability spaces $(\Omega, \mathcal{F}_n, P_n)$ are isomorphic, we can start considering what we mean by "the" unit interval with Lebesgue measure. Does there still exist a [illegible] entropy, they

cannot be isomorphic by 8.6.11. Thus the Bernoulli shift with distribution $(1/2, 1/2, 0)$ is not isomorphic to the Bernoulli shift with distribution $(1/4, 1/4, 1/4, 1/4)$. But what happens if two Bernoulli shifts have the same entropy? For example, two Bernoulli shifts were constructed in 4.2.1, 4.3.2, 4.3.3, and 4.3.4, one with entropy $\log 3$, and the other with entropy $\log 4$. Are they isomorphic? The answer is given in the following theorem which we state without proof.

8.7.4 *Theorem.* If two Bernoulli shifts have the same entropy, then they are isomorphic in the sense of 8.6.5.

8.8 References

The standard reference on measure theory is Halmos (1950). Some of the results of Section 8.2 may be found in Billingsley (1965). Some classical references on ergodic theory are Billingsley (1965), Halmos (1956), and Jacobs (1963). The pointwise ergodic theorem can be generalized in several ways. If T is a positive contraction operator on L^1, not necessarily arising from a measure-preserving transformation, one can investigate convergence of the sequence of averages $n^{-1}(f + Tf + \cdots + T^{n-1}f)$. In fact the arithmetic average can be replaced by a rather general form

$$a_0 f + a_1 T f + \cdots + a_{n-1} T^{n-1} f + a_n T^n f + \cdots + a_{n+k} T^{n+k} g,$$

where f and g belong to L^1. For results of this type (specifically, the Hopf-Dunford-Schwartz ergodic theorem and the Chacon-Ornstein theorem), see Garsia (1970).

A discussion of the Shannon-McMillan theorem for a finite or infinite space may be found in Billingsley (1965). McMillan (1953) proved the convergence in L^1, and Breiman (1957) obtained the convergence a.e. Shannon's original paper (1948) considered convergence in probability for functions of a finite Markov chain. All these results are for finite-dimensional spaces. The Shannon-McMillan theorem corresponds to Theorem 4 of Chung (1961). For applications to information theory, see Ash (1965) and Gallager (1968).

The Shannon-McMillan theorem has been generalized to a number of different settings. For a survey of this area and a unified approach to the various results, see Kingman (1968).

The definition of the entropy of a partition of a measure space is due to Kolmogorov (1958, 1959). The notion of Bernoulli shift can be generalized by allowing the coordinate space to have a countable number of factors. The entropy of such a shift can be defined. This provides the isomorphic

of Bernoulli shifts with source entropy in one to one correspondence with their Kolmogorov-Sinai entropy, and for generalized Bernoulli shifts with finite or infinite entropy (Ornstein, 1970b).

The Ornstein isomorphism theorem for Bernoulli shifts has been applied by Gray (1975) to show the equivalence of a class of sliding-block encoders (source codes) to a large class of ergodic sources.

CHAPTER

9

BROWNIAN MOTION AND STOCHASTIC INTEGRALS

9.1 STOCHASTIC PROCESSES

A *stochastic process* on a probability space $(\Omega, \mathcal{F}, P)$ is a family of random variables $\{X(t)\}_{t \in T}$, where the index set T can be any set. If $T = \mathbb{R}$, a stochastic process is simply a sequence of random variables X_n on the probability space $(\Omega, \mathcal{F}, P)$, and the values $X_n(\omega)$ may represent successive times for instance. A gambler plays at times $n = 1, 2, \ldots$; X_n is the total fortune of the gambler after X_n is the fortune at time n. In this chapter we will consider only the case $T = \mathbb{R}^+ = [0, \infty)$. Then the index t can again be thought as denoting time and for each fixed ω the function $t \mapsto X(t, \omega)$ is interpreted as the path of ω. We will see that the paths are continuous or that the process is continuous (resp. right-continuous) if the function $t \mapsto X(t, \omega)$ is continuous (resp. right-continuous) for each ω.

Stochastic processes are often constructed as mathematical models. For instance we can try to build a mathematical model for the number $N(t)$ of phone calls arriving at a switchboard in the interval of time $[0, t]$. Assume that the data support the following assumptions about the random variables N_t:

1. $N_0 = 0$.
2. If $0 \le t_1 < t_2 \le t_3 < t_4$, the increments $N_{t_2} - N_{t_1}$ and $N_{t_4} - N_{t_3}$ are independent.
3. The distribution of the increment $N_{t+h} - N_t$ has a Poisson distribution with parameter λh, $\lambda > 0$, where λ is the average number of calls in one unit of time.

One can construct the paths of the stochastic process $\{N_t\}_{t \ge 0}$ (called the Poisson process) in the following way. Let $T_1 \le T_2 \le \cdots \le T_n \le \cdots$ be the successive arrival times of the calls, and $W_1 = T_1$, $W_2 = T_2 - T_1$, $\ldots$, $W_n = T_n - T_{n-1}, \ldots$ the waiting times between calls. Assume that the W_n are independent identically distributed random variables with exponential distribution with mean $1/\lambda$. The process $N(t)$ is generated by $N_t = \sum_n I_{[0,t]}(T_n)$ as we did in Section 1.3. It is not very hard to see that for almost every ω, the path $t \mapsto$

$N_t(\omega)$ is a right-continuous, nondecreasing function that increases only by jumps of size 1. These are precisely the properties that we expect for the number of calls arriving in the interval $[0,t]$.

We can also use the Kolmogorov extension theorem to construct the Poisson process. Let $\Omega = \mathbb{R}^{[0,\infty)}$, the space of all functions $\omega : [0,\infty) \to \mathbb{R}$, denote by $\mathcal{B}$ the set $\{\omega : \omega(t_1) \in B_1, \ldots, \omega(t_n) \in B_n\}$ and by $\mathcal{F}_0$ the field of subsets of $\Omega = \mathbb{R}^{[0,\infty)}$ generated by the cylinder sets $X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n$. Let $\mathbb{P}_0$ be the probability on $(\Omega, \mathcal{F}_0)$ which makes $X_0 = 0$ a.s., and the $X_{t_i} - X_{t_{i-1}}$ independent and Poisson distributed with parameters $\lambda(t_i - t_{i-1})$. Set $\mathcal{F} = \mathcal{F}_\infty$. Then $(\Omega, \mathcal{F}, \mathbb{P}_0)$ is the sample space of a consistent system, thus we can apply the Kolmogorov extension theorem 2.2.3 and construct a probability $\mathbb{P}$ on $\mathcal{F}$ for which the X_t satisfy conditions 1–3. The problem is that we now have to study the paths $t \to X_t$. It is true that, outside of a set of $\mathbb{P}$-probability 0, the functions $t \to X_t$ are right-continuous and nondecreasing and increase only by jumps of size 1. (Unfortunately it is possible but not true (see Problem 2) that we have to modify each X_t on a set of probability 0 in such a way that the new paths satisfy all these qualitative properties. This will be the subject of Problem 3 in §4.)

The above discussion leads to the notion of version of a process. Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two stochastic processes on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The process $(Y_t)_{t\in T}$ is a *version* of the process $(X_t)_{t\in T}$ if for each $t \in T$ we have $\mathbb{P}[X_t = Y_t] = 1$. This does not imply the stronger condition $\mathbb{P}[X_t = Y_t \text{ for all } t \in T] = 1$. The set $\{X_t = Y_t \text{ for all } t \in T\}$ is generally not even measurable, but it does imply that for any countable subset J of T, $\mathbb{P}[X_t = Y_t \text{ for all } t \in J] = 1$. In particular the processes $(X_{q+})_{q\in\mathbb{Q}}$ and $(Y_q)_{q\in\mathbb{Q}}$ have the same finite-dimensional distributions. The regularity of the paths of a process can always be assumed in the weaker choice of version (see Problem 3).

In the next section we construct another stochastic process, Brownian motion $(B_t)_{t\in[0,\infty)}$. The random variables B_t will satisfy conditions 1 and 2 above, but, for most ω, the distribution of $B_t - B_s$ will be normal with mean 0 and variance $t - s$. Since the process is constructed as a model for the movement of a particle, we would like its paths to be continuous. Unfortunately there is no intuitive construction of Brownian motion that gives us continuous paths, so we will use Kolmogorov's extension theorem and then carefully choose a version with continuous paths.

We introduce the notion with a few easy consequences of just a regular σ-field. Let $(X_t)_{t\in T}$ be a stochastic process; since X_t is a random variable, for each fixed ω, the function $t \to X_t(\omega)$ is a function on the index set T to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The function $t \to X_t(\omega)$ is called the *path* of ω. We generally assume that all paths are measurable. But it is the case if the paths are all regular enough, for example if, for each ω, the function $t \to X_t(\omega)$ is right-continuous.

The stochastic process $(X_t)_{t\in T}$ can be considered as a function X of the two variables t and ω. The process $(X_t)_{t\in T}$ is said to be *measurable* if X is

mapping $\omega \mapsto X(t,\omega)$ is $\mathcal{F}$-measurable for each $t \in [0,\infty)$, then the process $\{X(t,\cdot)\}$ is measurable and, if X is not too large, we can consider integrals of the form $\int_0^t X(s,\omega)\,ds$ (Fubini's theorem). Again, continuity of the paths is sufficient to assure the measurability of X (see Problem 1).

EXERCISES

1. Let $\{X_t\}_{t\ge 0}$ be a stochastic process. Show that, if the paths are right-continuous (or left-continuous), the process $\{X_t\}_{t\ge 0}$ is measurable. (Hint: if the paths of the process are right-continuous, consider the processes X_t^n defined by $X_t^n(\omega) = X((k+1)/n,\omega)$ if $k/n \le t < (k+1)/n$, $k = 0, 1, \ldots$.)
2. Let $\{X_t\}_{t\ge 0}$ and $\{Y_t\}_{t\ge 0}$ be two processes on $(\Omega, \mathcal{F}, P)$. Assume that $\{Y_t\}_{t\ge 0}$ and $\{Y_t\}_{t\ge 0}$ are modifications of each other, i.e., $P\{X_t = Y_t\} = 1$ for all t, and that they both have right-continuous paths. Show that the set $\{\omega : X_t(\omega) = Y_t(\omega)\}$ for all rational t is measurable and has probability 1. (Show that the same holds if both processes have left-continuous paths.)
3. Let $\{X_t\}_{t\ge 0}$ be a continuous process. Assume that the given probability space has a random variable Z defined on it such that $P\{Z = x\} = 0$ for every real number x. By enlarging the probability space, show that there is a modification $\{Y_t\}_{t\ge 0}$ of $\{X_t\}_{t\ge 0}$ such that the paths of $\{Y_t\}_{t\ge 0}$ are nowhere continuous. (Hint: let $f(x) = 0$ if the function $t \to t - x$ is < 0 on the rationals and > 0 elsewhere. Consider the process $Y_t = X_t - f(Z + t)$.)

9.2 Brownian Motion

9.2.1 *Definition.* A *Brownian motion* is a stochastic process $\{B_t\}_{t\ge 0}$ with the following properties:

1. The increments on disjoint intervals are independent. If $0 < t_1 < t_2 < \cdots < t_n$, the random variables $B_{t_1} - B_0$, $B_{t_2} - B_{t_1}$, $\ldots$, $B_{t_n} - B_{t_{n-1}}$ are independent.
2. If $s < t$, the increment $B_t - B_s$ of the process on the interval $(s, t]$ is normally distributed with mean 0 and variance $t - s$.
3. The process starts at 0: $B_0 = 0$ with probability one.
4. The paths of the process B_t are all continuous.

9.2.2 *Remarks.* 1. The concept of Brownian motion goes back to the observations of the botanist Brown of the random movement of particles of pollen in water. Einstein and Smoluchowski showed that a good approximation for the projection of such a random movement on a line was given by the conditions in Definition 9.2.1, where the variance of $B_t - B_s$ is only assumed to be proportional to $t - s$. The constant of proportionality depends on the fluid. A rigorous proof of the existence of a continuous motion was given by Wiener.

For this reason the Brownian motion is often called the Wiener process and denoted by W.

2. The probability space $(\Omega, \mathcal{F}, P)$ is rich enough, and we can assume whatever seems appropriate.

3. Conditions 1, 2 and 3 involve only the finite dimensional distributions of the process. Condition 4 would $[W(t)]_{t\geq 0}$ is a process satisfying conditions 1, 2, 3, the increments $W_t - W_s = W_t - W_s$ is a Brownian motion. The existence of a process satisfying 1, 2, 3 will be an easy consequence of the Kolmogorov extension theorem. It will be much harder to construct a process satisfying condition 4 as well.

4. Condition 4 is sometimes weakened to "outside of a set of probability 0 the paths of $(X_t)_{t\geq 0}$ are continuous." Probabilistically, it does not matter whether we work with the stronger or the weak condition since we can always restrict ourselves to a set of probability 1 on which the paths are well behaved.

9.2.7 *Theorem.* There exists a probability P on the measurable space $(\mathbb{R}^{\mathbb{N}^*}, \mathcal{B}^{\mathbb{N}^*})$ such that for this probability, the coordinate functions X_t satisfy properties 1, 2, 3 of Definition 9.2.1.

Proof. We set $\Omega = \mathbb{R}^{\mathbb{N}^*}$ and $\mathcal{F} = \mathcal{B}^{\mathbb{N}^*}$. We denote by $\mathcal{E}$ the set $\{1, \ldots, n\}$ and by $\mathcal{P}(\mathcal{E})$ the set of subsets of $\Omega = \mathbb{R}^{\mathbb{N}^*}$ generated by the coordinates $X_1, \ldots, X_n$. Let $\pi_{\mathcal{E}}$ be the probability on $(\Omega, \mathcal{F})$ which makes $X_1 = \Pi_{\mathcal{E}}$ and $X_1, X_2 - X_1, \ldots, X_n - X_{n-1}$ independent and normally distributed with mean 0 and variances $\sigma_1 = t_1$, $\sigma_2 = t_2 - t_1, \ldots, \sigma_n = t_n - t_{n-1}$. If the $\pi_{\mathcal{E}}$ form a consistent system we can apply the Kolmogorov extension theorem 2.7.3, and the existence of probability P is proved. To verify the consistency of the $\pi_{\mathcal{E}}$, we have to show that if $\mathcal{E}$ denotes the union of $\mathcal{E}$ obtained by adding the element $n+1$, the restriction of $\pi_{\mathcal{E}}$ to the σ-field $\mathcal{F}_{\mathcal{E}}$ is $\pi_{\mathcal{E}}$. We know that for $\pi_{\mathcal{E}}$ the random variables $X_1 = X_1, X_2 - X_1, \ldots, X_n - X_{n-1}, X_{n+1} - X_n$ are independent and normally distributed with mean 0 and variances $\sigma_1 = t_1, \ldots, \sigma_n = t_n - t_{n-1}, \sigma_{n+1} = t_{n+1} - t_n$. We have to show that for $\pi_{\mathcal{E}}$ the random variables $X_1 = X_1, X_2 - X_1, \ldots, X_{n-1} - X_{n-2}, X_{n+1} - X_{n-1}, X_n - X_{n-1}$ are independent and normally distributed with mean 0 and variances $\sigma_1 = t_1, \ldots, \sigma_{n-1} = t_{n-1} - t_{n-2}, \sigma_n = t_{n+1} - t_{n-1}, \sigma_{n+1} = t_n - t_{n-1}$. This is true since the random variable $X_{n+1} - X_{n-1}$ is the sum of the two independent and normally distributed random variables $X_{n+1} - X_n$ and $X_n - X_{n-1}$. $\square$

9.2.8 *Remark.* If P is the probability constructed in 9.2.7, then, by Chebyshev's inequality, $P[|X_t - X_s| \geq \delta] \leq \frac{|t-s|}{\delta^2}$. The process X_t converges in probability to X_s as $t \to s$. For convergence in probability to be enough, we want pointwise convergence. Assume the existence of a continuous version

$(P_t)_{t\geq 0}$ of the process $(X_t)_{t\geq 0}$. Even if $(X_t)_{t\geq 0}$ has discontinuous paths, there will be a version which is continuous (or càdlàg) on the countable subsets of $\mathbb{R}^+$. If J is such a countable subset then the process $(X_t)_{t\in J}$ will be almost surely continuous on J, since the set $\{\omega : X_t(\omega) = t\}$ has probability one. Therefore our first step is to study the continuity of $(X_t)_{t\in J}$ on a countable dense subset of $\mathbb{R}^+$, such as the dyadics.

9.2.5 *Lemma.* Let $\Pi = \{t_0 < \cdots < t_n\}$. If the process $(X_t)_{t\geq 0}$ satisfies property (ii) of Definition 9.2.1, we have, for $\alpha > 0$,

$$P\left\{\max_{t\in\Pi} X_t \geq \alpha\right\} \leq 2P(X_{t_n} \geq \alpha)$$

and

$$P\left\{\max_{t\in\Pi} |X_t| \geq \alpha\right\} \leq 2P(|X_{t_n}| \geq \alpha).$$

Proof. Let A be the set $\{\max_{t\in\Pi} X_t \geq \alpha\}$ on the first time t_i such that $X_{t_i} \geq \alpha$. On the set $\{\max_{t\in\Pi} X_t \geq \alpha\}$ we have $T < \infty$. We have, using the fact that for $i = 0, \ldots, n-1$, the $X_{t_{i+1}} - X_{t_i}$ have symmetric distributions,

$$P\left\{\max_{t\in\Pi} X_t \geq \alpha\right\} = \sum_{i=0}^{n} P(T = t_i)$$

$$= \alpha \sum_{i=0}^{n} P(T = t_i) \cdot 2P(X_{t_n} - X_{t_i} \geq 0 \mid T = t_i).$$

Since the increment $X_{t_n} - X_{t_i}$ is independent of the σ-field $\sigma(X_{t_0}, \ldots, X_{t_i})$, which contains the set $\{T = t_i\}$, we obtain

$$\leq 2\sum_{i=0}^{n} P(T = t_i, X_{t_n} - X_{t_i} \geq 0) \leq 2\sum_{i=0}^{n} P(T = t_i, X_{t_n} \geq 0)$$

$$\leq 2P(X_{t_n} \geq 0).$$

For the second inequality, it is sufficient to remark that the process $(-X_t)_{t\geq 0}$ also satisfies condition (ii) of 9.2.1 and that

$$P\left\{\max_{t\in\Pi} |X_t| \geq \alpha\right\} \leq P\left\{\max_{t\in\Pi} X_t \geq \alpha\right\} + P\left\{\max_{t\in\Pi} -X_t \geq \alpha\right\}$$

$$\leq 2P(X_{t_n} \geq \alpha) + 2P(-X_{t_n} \geq \alpha).$$

9.2.6 *Corollary.* If the process $(X_t)_{t\geq 0}$ satisfies conditions 1–3 in 9.2.1, and J is a countable subset of $[0,\infty)$, then, for all $a > 0$, $P(\sup_{t\in J} X_t > a) = P(\sup_{t\in J} X_t \geq a)$.

9.2.7 *Theorem.* Let $(X_t)_{t\geq 0}$ be the process constructed in 9.2.3. For every $N > 0$ the restriction of the process $(X_{t,\omega})$ to the dyadic rationals is $\delta(N)$-uniformly continuous on $[0,N]$.

Proof. By rescaling the real line we can always assume that $N = 1$. Let $\mathbb{D}$ be the set of all dyadic rationals in $[0,1]$. We have to show that

$$Y_\delta(\omega) = \sup_{s,t\in\mathbb{D},\,|s-t|<\delta} |X_t(\omega) - X_s(\omega)| \to 0 \quad \text{as } \delta \to 0$$

and since $Y_\delta \geq Y_{\delta'}$ if $\delta \leq \delta'$ it is sufficient (see 9.2.1(c)) to show that for all $a > 0$, $P(Y_\delta > a) \to 0$.

The random variable Y_δ can be bounded above and controlled in terms of Z_δ by $2\max_k Z_{k,\delta}$ where

$$Z_{k,\delta} = \sup_{s\in\mathbb{D}\cap[k\delta,(k+1)\delta]} |X_s - X_{k\delta}| \quad k = 0, \ldots, \delta^{-1} - 1$$

and to use 9.2.6 to show that $P(\max_k Z_{k,\delta} > a) \to 0$ as $\delta \to 0$. As the random variables $X_{k\delta+s} - X_{k\delta}$, $0 \leq s \leq \delta$, are identically distributed,

$$P(\max_k Z_{k,\delta} > a) = P\left(\bigcup_k [Z_{k,\delta} > a]\right) \leq \sum_{k=0}^{\delta^{-1}-1} P(Z_{k,\delta} > a) = \delta^{-1} P(Z_{0,\delta} > a).$$

Now by 9.2.6,

$$P(Z_{0,\delta} > a) = P\left(\sup_{s\in\mathbb{D}\cap[0,\delta]} |X_s| > a\right) \leq 2P(|X_\delta| > a),$$

so that

$$P\left(\max_k Z_{k,\delta} > a\right) \leq 2\delta^{-1} P(|X_\delta| > a).$$

This last quantity converges to 0 as $\delta \to 0$ since X_δ has a normal distribution with mean 0 and variance δ^2 (see 9.2.8 below). ∎

9.2.8 *Lemma.* Let $a > 0$, $\beta > 0$ and assume that X_β is a normally distributed random variable with mean 0 and variance β. Then

$$\frac{1}{\beta} P(|X_\beta| > a) \to 0 \quad \text{as } \beta \to 0$$

Proof.

$$\frac{1}{s}P(|X_t| > s) = \frac{2}{t\sqrt{2\pi}s}\int_s^\infty e^{-\frac{x^2}{2t}}\,dx = \frac{2}{s\sqrt{\pi}}\int_{s/\sqrt{2t}}^\infty e^{-y^2}\,dy$$

$$< \frac{2}{s\sqrt{\pi}}\int_{s/\sqrt{2t}}^\infty \frac{y}{s/\sqrt{2t}}e^{-y^2}\,dy = \frac{\sqrt{2t}}{s\sqrt{\pi}}e^{-\frac{s^2}{2t}}.$$

If s is small enough to ensure that $s/\sqrt{2t} < 1$ $\square$

We are now ready to construct a continuous version of the process $(X_t)_{t\geq 0}$.

9.2.9 *Theorem.* There exists a process satisfying the conditions of 9.2.1 and 9.2.7.

Proof. With X_t defined as in the proof of 9.2.7, let A be the set $\{X_t = 0\}$. The set A is measurable and has probability 1. Define $Y_t(\omega)$ for all t and ω as the limit of the rational random variables $X_q(\omega)$ as q approaches t through the dyadic rationals (on the complement of A we take $Y_t(\omega) = 0$ for all t). By 9.2.7, the process $(Y_t)_{t\geq 0}$ is continuous, and we just have to prove that $(Y_t)_{t\geq 0}$ is a version of $(X_t)_{t\geq 0}$. Let $t \geq 0$, and (r_n) be a sequence of dyadic rationals converging to t. The random variables X_{r_n} converge almost surely and in probability to Y_t, which implies that $X_{r_n} \to Y_t$ in $\mathbb{L}^2$ $\square$

9.2.10 *Corollary.* If $(B_t)_{t\geq 0}$ is a Brownian motion, then the functions $\sup_{s\leq t} B_s$ and $\sup_{s\leq t} |B_s|$ are finite and satisfy, for all $\lambda > 0$, the following inequalities:

$$P(\sup_{s\leq t} B_s > \lambda) \leq 2P(B_t > \lambda) \quad \text{and}$$

$$P(\sup_{s\leq t} |B_s| > \lambda) \leq 2P(|B_t| > \lambda).$$

Proof. The continuity of the paths of $(B_t)_{t\geq 0}$ allows us to generalize 9.2.6 to $\mathbb{R}_+$ $\square$

Problems

1. Let $(X_t)_{t\geq 0}$ be a continuous process such that $(X_t)_{t\geq 0}$ is a Brownian motion. Show that for any finite subset $\{t_1, \ldots, t_n\}$ of $\mathbb{R}_+$, the random vector $(X_{t_1}, \ldots, X_{t_n})$ is Gaussian. What are its mean and covariance matrix? (See Appendix A for the definition and the properties of a Gaussian random vector.)

2. A sequence of processes $(X_t)_{t\geq 0}$ converges in probability to $(X_t)_{t\geq 0}$ if, for each $t \geq 0$, $X_t^n \to X_t$ in probability as $n \to \infty$. Show that, if two processes

$\{X_t : t \geq 0\}$ and $\{Y_t : t \geq 0\}$ have the same finite dimensional distributions, $\{X_t : t \geq 0\}$ has continuous sample paths if and only if $\{Y_t : t \geq 0\}$ has.

3. Let $\{B_t : t \geq 0\}$ be a Brownian motion and $c > 0$. Show that the process

$$X_t = \frac{1}{c} B_{c^2 t}$$

is a Brownian motion.

4. Let $\{B_t : t \geq 0\}$ be a Brownian motion. Show that the process

$$X_t = \begin{cases} t B_{1/t} & \text{if } t > 0, \\ 0 & \text{if } t = 0, \end{cases}$$

satisfies conditions 1–3 of Definition 5.2.1, and is continuous outside of a set of probability 0. The process $\{X_t : t \geq 0\}$ is therefore a Brownian motion and the properties of the Brownian motion can be used to prove [Hint: use Problem 3 to show that the process $\{X_t : t > 0\}$ satisfies conditions 1–3 of Definition 5.2.1 and then Theorem 5.2.2 to show that it is continuous at 0.]

5. Let $\{B_t : t \geq 0\}$ be a Brownian motion and $\{X_t : t \geq 0\}$ the process defined as

$$X_t = B_t - t B_1 \qquad 0 \leq t \leq 1.$$

The process $\{X_t : 0 \leq t \leq 1\}$ is called a Brownian bridge.

(a) Compute the means $E(X_t)$ and the covariances $E(X_s X_t)$. What is the distribution of X_t?

(b) Show that the process

$$Y_t = (1+t) B\left(\frac{t}{1+t}\right) \qquad 0 \leq t < \infty$$

has the same finite dimensional distributions as the process $\{B_t : t \geq 0\}$.

(c) Let $U_1, U_2, \ldots, U_n$ be independent random variables with uniform distributions on $[0, t]$ and let

$$F_n(t, \omega) = \frac{1}{n} \sum_{i=1}^{n} I_{[0,t]}(U_i(\omega)) \qquad 0 \leq t \leq 1$$

[$F_n(t, \omega)$ is an empirical distribution for the uniform distribution on $[0, 1]$.] Show that the process

$$X_n(t, \omega) = \sqrt{n}(F_n(t, \omega) - t), \qquad 0 \leq t \leq 1,$$

has the same means and covariances as the Brownian bridge.

(c) Show that for each t, the random variables X_t are converge in distribution as $n \to \infty$ to a random variable X_t.

5. Let $\{B_t\}_{t \geq 0}$ be a Brownian motion. Show that $B_t^2 - t$ is a martingale as follows.

(a) Let $X_1, X_2, \ldots$ be independent identically distributed random variables. Show that if $E|X_1| < \infty$, then for all $n > 0$, $P(\lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n} = 0) = 1$.

(b) Show that $B_t/n \to 0$ a.s. as the integer $n \to \infty$.

(c) Let

$$X_n = \max_{n \leq t \leq n+1} |B_t - B_n|$$

show that $E|X_n| < \infty$.

(d) Conclude by writing that for $n \leq t \leq n+1$

$$\left| \frac{B_t}{t} - \frac{B_n}{n} \right| \leq \frac{|B_t - B_n|}{n} + \left| \frac{B_n}{n} \right| \frac{X_n}{n}.$$

(e) Show that we could also used Problem 4 to prove directly that $B_t/t \to 0$ a.s. as $t \to \infty$.

[Hint: use the inequality $E(X) = \int_0^\infty P(X > \lambda)\, d\lambda$ for nonnegative random variables X, the strong law of large numbers and $0 \leq B_t^+$.]

7. The purpose of this exercise is to construct the Poisson process using the Kolmogorov extension theorem.

(a) Use the Kolmogorov Extension Theorem to construct a probability P on the measurable space $(\mathbb{R}^{\mathbb{R}^+}, \mathcal{B}^{\mathbb{R}^+})$, such that, for this probability, the coordinate functions X_t satisfy properties i, ii, iii below.

i. $X_0 = 0$ a.s.

ii. If $0 \leq t_1 < t_2 < \cdots < t_n$, the increments $X_{t_2} - X_{t_1}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent.

iii. For $s < t$, the increment $X_t - X_s$ has a Poisson distribution with parameter $\alpha(t-s)$.

(b) Let $\mathcal{E}$ be the set of dyadic *intervals*. Show that for each ω in a set of probability 0 we have

i. $X_0 = 0$

ii. The restriction of $\mathcal{E}$ to the values $\mathcal{E} \to V_\omega(\mathcal{E})$ are nondecreasing, and take only nonnegative integer values.

(c) Modify the process $(X_t)_{t \geq 0}$ to obtain a right-continuous, integer valued process $(N_t)_{t \geq 0}$ satisfying the properties in question 6(a) and whose paths are nondecreasing.

(d) Show that, outside of a set of probability 0, the paths of $(N_t)_{t \geq 0}$ increase only by jumps of size 1.

[Hint: Let $Y = 0$ be those ω for which the following limit does not exist: $\lim_{\varepsilon \to 0} \sup_{0 \leq t \leq 1} |N_{t+\varepsilon} - N_t| = 0$.]

9.3 Nowhere Differentiability and Quadratic Variation of Paths

The sample paths of Brownian motion are continuous but they are nowhere differentiable.

9.3.1 *Theorem.* Let $(B_t)_{t\geq 0}$ be a Brownian motion on a probability space $(\Omega, \mathcal{F}, P)$. Then the paths of $(B_t)_{t\geq 0}$ are nowhere differentiable.

As stated above, the theorem is slightly misleading. The set $\{\omega : t \mapsto B_t(\omega)$ is differentiable at some point$\}$ is not necessarily measurable, but we will show that it is included in a measurable set of probability 0.

Proof.

We start by studying what differentiability at one point implies for a path. Let $f(t)$ be a real-valued function on $[0,\infty)$, assume that f has a derivative $f'(t_0)$ at a point t_0, and choose a real number $A > |f'(t_0)|$. Then there exists a positive integer n_0 such that for every $n \geq n_0$,

$$|f(t) - f(s)| \leq 2A\,|t - s| \qquad \text{if } |t - t_0| \leq 2/n. \tag{1}$$

For $n \geq n_0$, let k be the integer such that $k/n \leq t_0 < (k+1)/n$. The points $(k-1)/n$, k/n, $(k+1)/n$, $(k+2)/n$ are within distance $3/n$ of t_0, so applying (1) and the triangle inequality we obtain, for example,

$$\left|f\left(\frac{k+2}{n}\right) - f\left(\frac{k}{n}\right)\right| \leq \left|f\left(\frac{k+2}{n}\right) - f(t_0)\right| + \left|f(t_0) - f\left(\frac{k+1}{n}\right)\right| + \left|f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right)\right| \leq 2A\left|\frac{k+2}{n} - t_0\right| + 2A\left|\frac{k+1}{n} - t_0\right| + \frac{2A}{n} \leq \frac{10A}{n}.$$

Using the same method, we obtain similar inequalities for $|f(k/n) - f((k-1)/n)|$ and $|f((k-1)/n) - f((k-2)/n)|$, which gives us

$$\max\left\{\left|f\left(\frac{k-2}{n}\right) - f\left(\frac{k-1}{n}\right)\right|, \left|f\left(\frac{k-1}{n}\right) - f\left(\frac{k}{n}\right)\right|, \left|f\left(\frac{k}{n}\right) - f\left(\frac{k+1}{n}\right)\right|, \left|f\left(\frac{k+1}{n}\right) - f\left(\frac{k+2}{n}\right)\right|\right\} \leq \frac{10A}{n}.$$

Now we apply all this to the Brownian motion. Denote by X_n the random variable

$$X_n = \min_{0\leq k\leq n-2}\left\{\left|B\left(\frac{k-1}{n}\right) - B\left(\frac{k}{n}\right)\right|, \left|B\left(\frac{k}{n}\right) - B\left(\frac{k+1}{n}\right)\right|, \left|B\left(\frac{k+1}{n}\right) - B\left(\frac{k+2}{n}\right)\right|\right\}.$$

Let A be the set of ω's such that, somewhere on $[0,1]$, the function $t \mapsto B_t(\omega)$ has a finite derivative in absolute value by a constant A. According to our

preceding discussion, we see that if n is large enough, then for each k in the set A_n at least one of the four inequalities $|X(\frac{k+i}{n}) - X(\frac{k+i-1}{n})| \le 4M/n$, $i = 0, 1, 2, 3$, will hold. If we show that $P(\limsup_n A_n) = 0$, we will have proved that $P\{\omega : D^+ f(t,\omega) = \infty\} = 0$. The other four derivatives can be dealt with in a similar way.

Since the increments of the Brownian motion on the intervals $[j/n, (j+1)/n]$ are independent and identically distributed, we have

$$\begin{aligned} P(A_n) &\le \sum_{k=1}^{n} P\left\{ |X_{\frac{k+1}{n}} - X_{\frac{k}{n}}| \le \frac{12M}{n}, \right. \\ &\qquad \left. |X_{\frac{k+2}{n}} - X_{\frac{k+1}{n}}| \le \frac{12M}{n}, \; |X_{\frac{k+3}{n}} - X_{\frac{k+2}{n}}| \le \frac{12M}{n} \right\} \\ &= n P\left\{ |X_{\frac{1}{n}} - X_0| \le \frac{12M}{n}, \; |X_{\frac{2}{n}} - X_{\frac{1}{n}}| \le \frac{12M}{n}, \right. \\ &\qquad \left. |X_{\frac{3}{n}} - X_{\frac{2}{n}}| \le \frac{12M}{n} \right\} \\ &= n \left| P\left\{ |X_{\frac{1}{n}} - X_0| \le \frac{12M}{n} \right\} \right|^3 \\ &= n \left[\frac{1}{\sqrt{2\pi/n}} \int_{-12M/n}^{12M/n} e^{-x^2/(2/n)} \, dx \right]^3 \\ &= n \left| \frac{n}{\sqrt{2\pi}} \int_{-12M/n}^{12M/n} e^{-nx^2/2} \, dx \right|^3 \to 0. \end{aligned}$$

We have been studying differentiability on the interval $[0, 1]$; what the bound is over the same technique works on $[0, N]$ with the inequalities changed slightly. The theorem is proved. $\square$

9.3.2 *Remark.* Theorem 9.3.1 can be generalized to upper, right and left derivatives. The proof can be modified to show that, outside of a set of probability 0, $\limsup_{h \to 0} (B_{t+h} - B_t)/h = +\infty$, $\liminf_{h \to 0} (B_{t+h} - B_t)/h = -\infty$, $\limsup_{h \to 0} (B_{t+h} - B_t)/h = +\infty$, and $\liminf_{h \to 0} (B_{t+h} - B_t)/h = -\infty$ are nowhere finite.

9.3.3 *Corollary.* Almost every path of the Brownian motion has infinite variation on every finite interval.

Proof. By 2.3.9, if a path has finite variation on an interval $[a, b]$, then it is almost everywhere differentiable on the interval. $\square$

If $[a, b]$ is an interval of length $h = (b-a)/n$ and $V = \{a, a+h, \ldots, b\}$ is any partition of $[a, b]$, we have just shown that $\sum_{t \in V} |B(t) - B(t-h)| \to \infty$. What can we say about the quadratic variation $\sum_{t \in V} |B(t) - B(t-h)|^2$ when the partitions are fine?

9.14 *Theorem.* Suppose $\sup_x |G_n(x) - G(x)| \xrightarrow{a.s.} 0$, where G is a continuous cdf and $\mathcal{P}_n = \{\xi_0^{(n)}, \ldots, \xi_n^{(n)}\}$ is a sequence of partitions of $[a, b]$. Then

$$\sum_{j=1}^{n} \hat{R}(\xi_j^{(n)}) \left[\hat{S}(\xi_j^{(n)}) - \hat{S}(\xi_{j-1}^{(n)}) \right]^2 \xrightarrow{p} \psi \quad \text{as} \quad n \to \infty \quad \text{and} \quad \max_j \left| \xi_j^{(n)} - \xi_{j-1}^{(n)} \right| \to 0.$$

Proof. Let $S_n = \sum_{j=1}^{n} \hat{R}(\xi_j^{(n)}) \left[\hat{S}(\xi_j^{(n)}) - \hat{S}(\xi_{j-1}^{(n)}) \right]^2$; we have

$$\begin{aligned} E(S_n - \psi - a)^2 &= E\Big(\sum_j \big[R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)}) \big]^2 \big[S(\xi_j^{(n)}) - S(\xi_{j-1}^{(n)}) \big]^2 \\ &\quad - \sum_j \big[R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)}) \big] \big[S(\xi_j^{(n)}) - S(\xi_{j-1}^{(n)}) \big] \\ &\qquad \big[R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)}) \big] \big[S(\xi_j^{(n)}) - S(\xi_{j-1}^{(n)}) \big] \Big)^2 \\ &= E \sum_j \big[R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)}) \big]^2 \big[S(\xi_j^{(n)}) - S(\xi_{j-1}^{(n)}) \big]^2. \end{aligned}$$

since $\sum_j [R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)})] [S(\xi_j^{(n)}) - S(\xi_{j-1}^{(n)})]$ are independent mean zero variables with mean 0. The random variables $R(\xi_j^{(n)}) - R(\xi_{j-1}^{(n)})$ are normally distributed with variance $\frac{\xi_j^{(n)} - \xi_{j-1}^{(n)}}{n}$. Therefore, if Z denotes a standard normal random variable with mean 0 and variance 1,

$$\begin{aligned} n E(S_n - \psi - a)^2 &= n E\Big[\big(\tfrac{Z}{\sqrt{n}} \big)^2 \big(\tfrac{Z}{\sqrt{n}} \big)^2 \Big] = \frac{1}{n} E[Z^4] \sum_j (\xi_j^{(n)} - \xi_{j-1}^{(n)})^2 \\ &\le \frac{3}{n} \max_j \left| \xi_j^{(n)} - \xi_{j-1}^{(n)} \right| \to 0. \end{aligned}$$

Problems

1. Show that if the Bernoulli cdf has a slope greater than one over the interval $[0, 1]$, then the process A_n given in Problem 9.13 has a limiting distribution which is not Brownian motion. This shows intuitively why the Bernstein statistic is not nonlinear differentiable.

2. Let ψ_n and S_n be defined as in 9.13 and set $|\mathcal{P}_n| = \max_j |\xi_j^{(n)} - \xi_{j-1}^{(n)}|$. Show that if for a sequence of partitions $\mathcal{P}_n$, $\sum_n |\mathcal{P}_n|^2 < \infty$, then $S_n \to \psi$ a.s.

9.4 Kolmogorov-Smirnov Tests

In $(R(\xi_j), S(\xi_j))$ Brownian motion, we know that $R(\xi_j) \to R(\xi)$ as $j \to \infty$ (problem 6 of 9.2). In Section 9.5, we show that more precisely $R(\xi_j)$ converges asymptotically to between $\sqrt{2n \log\log(n)}$ and $\sqrt{4n \log\log(n)}$ as $n \to \infty$. Assuming ψ

Problem 4 of §9.2, the behavior of B_t at 0 and ∞ are related, and it is not surprising that a similar asymptotic property is satisfied at 0 (see Theorem 9.4.3).

The following estimate will be useful in the proof of (a) $\Rightarrow$ (b).

9.4.1 *Lemma.* Let $X_1, X_2, \ldots$ be normally distributed random variables with mean 0 and variance 1, and let a be strictly positive. Then

$$P\{X_j > a\sqrt{j}\} \geq \frac{a^2}{a^2+2\pi} e^{-a^2 j/2}.$$

Proof. Just notice that

$$\begin{aligned} P\{X_j > a\sqrt{j}\} &= \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{j}}^{\infty} e^{-x^2/2}\,dx \\ &\geq \frac{1}{a\sqrt{2\pi}} \int_{a\sqrt{j}}^{\infty} x e^{-x^2/2}\,dx \\ &= \frac{\sqrt{2}}{a\sqrt{2\pi}} e^{-a^2 j/2}. \end{aligned}$$

9.4.2 *Lemma.* The events $\{B_{2^n} > a\sqrt{2^n}\}$ and $\{B_{2^n} > a\sqrt{2^{n-1}}\}$ are convergent to 1 or convergent to 0 (this is the inequality of §9.4 as an asymptotic equality as $n \to \infty$).

Proof. Apply the Borel–Cantelli Lemma. ∎

9.4.3 *Theorem.* If B_t is a Brownian motion, then we have

$$\limsup_{t \to 0} \frac{B_t}{\sqrt{2t \log\log(1/t)}} = 1 \quad \text{a.s.}$$

and

$$\liminf_{t \to 0} \frac{B_t}{\sqrt{2t \log\log(1/t)}} = -1 \quad \text{a.s.}$$

Proof. The second statement of the theorem is a consequence of the first because $-B_t$ is also a Brownian motion.

I. Let $\psi(t) = \sqrt{2t \log\log(1/t)}$. We first show that

$$\limsup_{t \to 0} \frac{B_t}{\psi(t)} \leq 1 \quad \text{a.s.}$$

This is equivalent to: for almost every ω, and for every $\delta > 0$, $B_t(\omega) \leq (1+\delta)\psi(t)$ when t is sufficiently small.

Choose $\theta \in (0,1)$, choose $\alpha \in (1, \frac{1+\theta}{1-\theta})$ and let $a_n = (1+\theta)\sqrt{2\log\log(\alpha^n)}$; consider the non-decreasing sequence of sets A_n in the set

$$A_n = \{\omega : B_t(\omega) > a_n \sqrt{t(1+\theta)} \text{ for at least one } t \in [0, \alpha^{n-1}]\}$$

Using 9.2.10 and the fact that the function $\omega(t)$ is increasing for sufficiently small t, we obtain

$$\begin{aligned} P(A_n) &= P(|u_n| \le \delta_n \sqrt{2n \log\log n}) \\ &\le 2P(B_n) = (1 - \Phi(\delta_n)) + 1. \end{aligned}$$

Therefore, by (9.2.1),

$$P(A_n) \le \left(\frac{2}{\pi}\right)^{1/2} \frac{e^{-\delta_n^2/2}}{\delta_n (\log n)^{1/2}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Substituting $\delta_n^2 = 2\alpha \log\log n$, we obtain for example

$$\begin{aligned} P(A_n) &\le \left(\frac{2}{\pi}\right)^{1/2} \frac{1}{(2\alpha \log\log n)^{1/2} (\log n)^{\alpha}} + o\left(\frac{1}{\sqrt{n}}\right) \\ &< A \frac{1}{(n+1)^{\alpha} \sqrt{\log\log n + 1}} \end{aligned}$$

where $\alpha = (1 - \varepsilon)^2$ and $A > 1$. Therefore $\sum P(A_n) < \infty$, and, by the first Borel–Cantelli lemma (9.4.1), $P(\limsup_n A_n) = 0$. This means that, for almost every ω, $|S_n(\omega)| \le \delta_n \sqrt{2n \log\log n}$ when n is sufficiently large.

(b) We now show that $\limsup_{n \to \infty} |S_n(\omega)| / \sqrt{2n \log\log n} \ge 1$ a.e. by showing that, for any $\varepsilon > 0$, there exists a sequence $n_1 < n_2 < \cdots$ of integers such that $P(B_{n_k} \text{ i.o.}) = 1$, where $B_n = \{|S_n| \ge (1 - \varepsilon)\sqrt{2n \log\log n}\}$. (Recall that the abbreviation "i.o." stands for "infinitely often".)

We now choose $\alpha > 0$ and $\beta \in (0, 1)$ small enough to guarantee first $(1 - \beta)^2 > (1 - \varepsilon)^2$, and $(1 - \alpha)(1 + \beta)(1 - \beta) > 1 - \beta$. For $n_k = 2^k$, we consider the independent random variables $X_k = S_{n_k} - S_{n_{k-1}}$. We want to estimate $\sum P(X_k \ge (1 - \varepsilon)\sigma_k)$.

$$\begin{aligned} P(X_k \ge (1 - \varepsilon)\sigma_k) &= P\left(\frac{X_k}{\sqrt{n_k - n_{k-1}}} \ge \frac{(1 - \varepsilon)\sigma_k}{\sqrt{n_k - n_{k-1}}}\right) \\ &= P\left(\frac{X_k}{\sqrt{n_k - n_{k-1}}} \ge \frac{(1 - \varepsilon)\sqrt{2\alpha \log\log n_k}}{\sqrt{1 - \alpha}}\right) \end{aligned}$$

Therefore, applying Lemma 9.4.2, we get, since $A_n / \sqrt{n} \to 1$, that $\bar{X}_n$ has a normal distribution with mean 0 and variance 1.

$$P(X_n > 0) \quad \text{[illegible]} \quad \frac{K}{n^{\alpha} \sqrt{\log n}}$$

where $\alpha = 1 - p/2 = p/2 + \ldots$ [illegible] Since $\sum_n P(X_n > 0) = \infty$, but, by the second Borel-Cantelli Lemma 6.3.2 we have

$$P(X_n > 0 \text{ i.o.}) = P(\limsup_n \{X_n > 0\}) = 1$$

If we apply part 1 of the proof to the case $-X_n$, we see that

$$P(S_{n+1} \geq \quad [1 + o(1)] \sqrt{n \log n}) \text{ for a large enough } n) = 1,$$

and therefore,

$$P(S_n \geq \alpha \sqrt{n}) \left[1 - \Phi(\quad + o(1) + \alpha \frac{\sqrt{n \log n}}{\sqrt{n}} \right] \text{ i.o.}) = 1.$$

Since $\min_{\alpha > 0} [\alpha + \alpha^2/2] = -1/2$ we obtain

$$P(S_n > (1 - \delta) \sqrt{n \log n} \text{ i.o.}) = 1$$

We give now the [illegible] We could prove the theorem directly, but it is easier to use Problem 9.6. If S_n is a Brownian motion, so is the process $X_t = \sigma B_{t/\sigma^2}$, and the properties of the paths of the Brownian motion will be preserved.

9.4.4 *Theorem* If S_t is a Brownian motion, then we have

$$\limsup_{t \to \infty} \frac{S_t}{\sqrt{2t \log \log t}} = 1 \quad \text{a.s.}$$

and

$$\liminf_{t \to \infty} \frac{S_t}{\sqrt{2t \log \log t}} = -1 \quad \text{a.s.}$$

Proof Apply Theorem 9.4.3 to the process $X_t = \sigma B_{t/\sigma^2}$ and take $\sigma = 1/2$.

Problems

1. Let $X_n = \sum_{k=1}^{n} Y_k$, where the Y_k are independent and each has a normal distribution with mean 0 and variance σ^2. Show that

$$\limsup_{n \to \infty} \frac{X_n}{\sqrt{2n \log \log n}} = \sigma \quad \text{a.s.}$$

and

$$\inf\left\{\frac{\cdot}{\sqrt{\cdot}\log\log(\cdot)}=-1\right\}=\cdot$$

9.5 The Markov Property

The Markov property discussed in Chapter 6 can be expressed as follows: a process $(X_t)_{t\geq 0}$ satisfies the Markov property if the prediction of X_t knowing what happened up to time s (for $s<t$) depends on the value of X_s and not on the values of X_u for $u<s$. The Markov property of the Brownian motion is a corollary of the definition of the Brownian motion.

9.5.1 Theorem (Markov Property). Let $(B_t)_{t\geq 0}$ be a Brownian motion. Then for every $t\geq 0$, the process $Y_s = B_{t+s} - B_t$, $s\geq 0$, is a Brownian motion independent of the σ-field $\mathcal{F}_t = \sigma(B_s,\ s\leq t)$.

We would like to generalize this property by allowing t to be a stopping time. We give the definition of stopping times with respect to a general non-decreasing family of σ-fields since it will be needed in the problems, but in this section, we are only interested in the family $(\mathcal{F}_t)_{t\geq 0}$.

9.5.2 Definition. Let $(\Omega,\mathcal{F},P)$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ be a non-decreasing family of sub-σ-fields of $\mathcal{F}$. A non-negative random variable T is a stopping time for the σ-fields $(\mathcal{F}_t)_{t\geq 0}$ if for each $t\geq 0$, the event $\{T\leq t\}$ belongs to $\mathcal{F}_t$. The random variable T is allowed to take the value $+\infty$.

The σ-field $\mathcal{F}_T$ of events prior to T is

$$\mathcal{F}_T = \{A\in\mathcal{F} : A\cap\{T\leq t\}\in\mathcal{F}_t,\ \forall t\geq 0\}.$$

It is also useful to consider the σ-field

$$\mathcal{F}_{T-} = \sigma\{A\cap\{T>t\} : A\in\mathcal{F}_t,\ t\geq 0\}.$$

These definitions are consistent with the definitions given in Chapter 6 since for the jump set $\{\Delta\neq 0\}$, the conditions $\{\Delta=a\}\in\mathcal{F}_a$ for any $a>0$, and $\{\Delta=a\}\in\mathcal{F}_{a-}$ for any $a>0$ are equivalent. When the index set is $\mathbb{N}$, the two conditions are no longer equivalent, and the condition $\{T=n\}\in\mathcal{F}_n$ for any n is much weaker than the more natural measurable requirement such as $\{T\leq n\}=\bigcup_{k\leq n}\{T=k\}$.

Constant times $T=t$ are stopping times in this case. $\mathcal{F}_{t-}=\mathcal{F}_t$ and $\mathcal{F}_{t+}=\mathcal{F}_t=\bigcap_{s>t}\mathcal{F}_s$.

The following lemma gives some useful properties of stopping times and their associated σ-fields.

9.5.3 Theorem. Let $(\Omega,\mathcal{F},P)$ be a probability space, $(\mathcal{F}_t)_{t\geq 0}$ a non-decreasing family of sub-σ-fields of $\mathcal{F}$, and S and T be two stopping times for the family $(\mathcal{F}_t)_{t\geq 0}$.

(a) We have the inclusion $\mathcal{F}_{S\wedge T}\subset\mathcal{F}_S$.

(3.6) The random variable T is $\mathcal{F}_T$-measurable.
(3.7) If $S \le T$, then $\mathcal{F}_S \subset \mathcal{F}_T$ and $\mathcal{F}_{S-} \subset \mathcal{F}_{T-}$.
(3.8) If $S \le T$, then $\mathcal{F}_{S-} \subset \mathcal{F}_T$.
(3.9) If Z is an $\mathcal{F}_T$-measurable random variable and $S \le T$, then Z is a $\mathcal{F}_S$-measurable.
(3.10) If S and T are two stopping times, then $S \vee T$ and $S \wedge T$ are stopping times and $\mathcal{F}_{S \wedge T} = \mathcal{F}_S \cap \mathcal{F}_T$.
(3.11) If S and T are two stopping times, then $S + T$ is also a stopping time.

Proof.

(3.6) Let $t \in [0, \infty]$. Then for any $r \in \mathbb{Q}$ the set $\{T \le t\} \cap \{T \le r\} = \{T \le r\}$ is in $\mathcal{F}_r$ and $r \le t$, so $\{T \le t\} \in \mathcal{F}_t$.

(3.7) It is necessary to show that, for any $t \ge 0$, the set $\{S \le t\}$ is in $\mathcal{F}_T$ for any $t \ge 0$. We have

$$\{S \le t\} \cap \{T \le t\} = \{S \le t \wedge T\} \in \mathcal{F}_{t \wedge T}.$$

(3.8) Assume that $A \in \mathcal{F}_S$. Then $A \cap \{T \le t\} = A \cap \{S \le t\} \cap \{T \le t\} \in \mathcal{F}_t$, and $A \in \mathcal{F}_T$. Similarly for the σ-fields $\mathcal{F}_{S-}$ and $\mathcal{F}_{T-}$.

(3.9) Use the equality $A \cap \{T < t\} = \bigcup_{r \in \mathbb{Q}} A \cap \{S \le r\} \cap \{T < t\} \cap \{T \le r\}$ to conclude.

(3.10) Use the equalities $\{T \le t\} = \{S \le t\} \cap \{T \le t\}$ to conclude.

(3.11) To show that $S \vee T$ is a stopping time, use the equality $\{S \vee T \le t\} = \{S \le t\} \cap \{T \le t\}$. Similarly $\{S \wedge T \le t\} = \{S \le t\} \cup \{T \le t\}$. Hence we see that $\mathcal{F}_{S \wedge T} \subset \mathcal{F}_S \cap \mathcal{F}_T$. Conversely, if $A \in \mathcal{F}_S \cap \mathcal{F}_T$, we have $A \cap \{S \wedge T \le t\} = (A \cap \{S \le t\}) \cup (A \cap \{T \le t\}) \in \mathcal{F}_t$.

(3.12) Since S is $\mathcal{F}_S$-measurable, the sets $\{S \le t\} \cap \{T \le t\}$ and $\{S \le t\} \cap \{T \le t\}$ are in $\mathcal{F}_t$; the same set is in $\mathcal{F}_t$ which means that the restriction of S to the set $\{S \le t\} \cap \{T \le t\}$ is $\mathcal{F}_t$-measurable. A similar property holds for T. Therefore the restriction of $S - T$ to the set $\{S \le t\} \cap \{T \le t\}$ is $\mathcal{F}_t$-measurable and the set $\{S - T \le r\} \cap \{S \le t\} \cap \{T \le t\}$ is in $\mathcal{F}_t$. □

3.14 Remark. It is often assumed in the general theory of stochastic processes that the family of σ-fields $\{\mathcal{F}_t\}$ is right-continuous: $\mathcal{F}_t = \bigcap_{s > t} \mathcal{F}_s$. In any case $\{\mathcal{F}_t\}$ is right-continuous in probability: $\mathcal{F}_t = \mathcal{F}_{t+}$ for any t is equivalent to $\{T < t\} \in \mathcal{F}_t$ for any t, and the σ-fields $\mathcal{F}_T$ and $\mathcal{F}_{T+}$ are the same. The usually considered natural σ-fields satisfy the Blumenthal–Getoor–Meyer right-continuous, but we will see in Problem 2.7 that $\mathcal{F}_t$ and $\mathcal{F}_{t+}$ differ only by measurable sets of probability 0.

The following lemma is useful to verify that hitting times are stopping times.

9.5.7 **Lemma.** Let $g(t)$ be a continuous function on $[0,\infty)$ such that $g(0) = 0$, $\limsup_{t\to\infty} g(t) = \infty$, and $T = \inf\{t > 0 : g(t) = a\}$ for some $a > 0$. Then $T < \infty$ a.s. if and only if $\sup_{0\le t\le s} g(t) \ge a$ for all $s > 0$. (A similar result holds for $a < 0$.)

Proof. As the function g is continuous, $\sup_{0\le t\le s} g(t) = \sup_{0\le t<s} g(t)$ for any $g(T) = a$. If $T < \infty$, then $\sup_{0\le t\le T} g(t) \ge g(T) = a$. Conversely, if $\sup_{0\le t\le s} g(t) \ge a$ for all $s > 0$, then there exists a point t_0 in $[0,1]$ such that $g(t_0) = \sup_{0\le t\le 1} g(t) \ge a$. Hence g reaches a at some point in $[0,1]$, and therefore $T \le 1$. □

9.5.8 **Example.** Let $\{B_t\}_{t\ge 0}$ be a Brownian motion and $a \ne 0$. The first time B_t reaches a is usually denoted as $T_a = \inf\{t > 0 : B_t = a\}$. The Law of the Iterated Logarithm ensures that the time T_a is almost surely finite, and the continuity of the Brownian motion ensures that $B_{T_a} = a$ on $\{T_a < \infty\}$. For $a > 0$, then $T_a > t$ if and only if $\sup_{0\le s\le t} B_s < a$, which shows that T_a is a stopping time for the family $\{\mathcal{F}_t^B\}_{t\ge 0}$. A similar argument gives the result for $a < 0$.

We want to generalize the Markov property to stopping times, and we begin with a result on the measurability of B_T.

9.5.9 **Theorem.** Let $\{B_t\}_{t\ge 0}$ be a Brownian motion and T be a finite stopping time for the family $\{\mathcal{F}_t^B\}_{t\ge 0}$. Then B_T is $\mathcal{F}_T^B$-measurable.

Proof. Define $X_n = B_{k/2^n}$ on $\{(k-1)/2^n \le T < k/2^n\}$, $k = 0, 1, 2, \ldots$. Then X_n is $\mathcal{F}_T^B$-measurable, and $B_T = \lim_{n\to\infty} X_n$. □

9.5.10 **(Strong Markov Property).** Let $\{B_t\}_{t\ge 0}$ be a Brownian motion and T be a finite stopping time for the family $\{\mathcal{F}_t^B\}_{t\ge 0}$. Then $\{B_{T+t} - B_T\}_{t\ge 0}$ is a Brownian motion independent of the σ-field $\mathcal{F}_T^B$.

Proof. Consider the times T_n defined as follows:

$$T_n = \frac{k}{2^n} \quad \text{if} \quad \frac{k-1}{2^n} \le T < \frac{k}{2^n}, \; k = 1, 2, \ldots.$$

The T_n are stopping times (9.5.2(iv)), and $A \subset \mathcal{F}_T^B$ implies that $A \cap \{T_n = k/2^n\}$ is in $\mathcal{F}_{k/2^n}^B$. Applying 9.5.1, we get, for $A \subset \mathcal{F}_T^B$,

$$\begin{aligned} P(\{B_{T_n+t} - B_{T_n} \le x\} \cap A) &= \sum_{k=1}^{\infty} P(\{B_{(k/2^n)+t} - B_{k/2^n} \le x\} \cap A \cap \{T_n = k/2^n\}) \\ &= \sum_{k=1}^{\infty} P(B_{k/2^n+t} - B_{k/2^n} \le x) P(A \cap \{T_n = k/2^n\}) \\ &= P(B_t \le x) P(A). \end{aligned}$$

We now restrict the set A to be open instead. Let $n = 1$. When $n = 1$, the random variables $B_{T_k} - B_{T_{k-1}}$ are independent and identically distributed. Moreover $P(|B_{T_k} - B_{T_{k-1}}| \le \varepsilon) = P(|B_{T_1}| \le \varepsilon) < 1$ for $\varepsilon > 0$. Hence $P(|B_{T_k} - B_{T_{k-1}}| \le \varepsilon) \le \delta < 1$ for some δ and all k. Therefore $P(|B_{T_k} - B_{T_{k-1}}| \le \varepsilon) \le \delta^k$. Letting k approach infinity, we get $P(|B_{T_k} - B_{T_{k-1}}| \le \varepsilon) = 0$. Since $\varepsilon > 0$ is arbitrary, we get $P(B_{T_k} - B_{T_{k-1}} = 0) = 1$. Hence $P(B_{T_k} = B_{T_{k-1}}) = 1$. Therefore $P(B_{T_k} = B_{T_{k-1}} = \cdots = B_{T_1} = B_0) = 1$. Letting k approach infinity, we get $P(B_t = B_0 \text{ for all } t \in [0, \infty)) = 1$. Hence $P(B_t = B_0 \text{ for all } t \in [0, \infty)) = 1$. Letting $B_0 = b$, we get $P(B_t = b \text{ for all } t \in [0, \infty)) = 1$.

$$P(|B_{t_1} - B_0| + \cdots + |B_{t_n} - B_{t_{n-1}}| \le \varepsilon) = P(|B_{t_1}| + \cdots + |B_{t_n} - B_{t_{n-1}}| \le \varepsilon) \tag{1}$$

For any $\varepsilon > 0$, we have since we have already established the Strong Markov Property for the stopping times T_k.

$$P(|B_{T_1}| + \cdots + |B_{T_n} - B_{T_{n-1}}| \le \varepsilon) = \lim_{k \to \infty} P(|B_{T_1}| + \cdots + |B_{T_n} - B_{T_{n-1}}| \le \varepsilon)$$
$$= P(|B_{t_1}| + \cdots + |B_{t_n} - B_{t_{n-1}}| \le \varepsilon)$$

Let $\mathcal{F}$ be the σ-field $\mathcal{F}(B_{T_1}, \ldots, B_{T_n})$ for any $n \ge 0$. Therefore equality (1) is satisfied for all $\varepsilon > 0$ and $B_{T_1}, \ldots, B_{T_n}$ is a Brownian motion independent of $\mathcal{F}_{T_n}$. $\square$

9.5.9 *Remark.* The same proof shows that, more generally, if T is a finite stopping time for a family of σ-fields $\mathcal{F}_t$, such that $B_{T+s} - B_T$ is a Brownian motion independent of the σ-field $\mathcal{F}_T$, the stopping times T_k defined as above are still stopping times for the family $(\mathcal{F}_t)_{t \ge 0}$ and as $T = T_k$, we have $\mathcal{F}_T \subset \mathcal{F}_{T_k}$. Therefore the above proof is still valid if we replace $\mathcal{F}_T$ by $\mathcal{F}_{T_k}$.

In particular, if T is a finite stopping time for the family $(\mathcal{F}_t)_{t \ge 0}$, then $B_T - B_0$ is a Brownian motion independent of the σ-field $\mathcal{F}_T$.

We now apply the Strong Markov property to the problem of finding the distribution of the random variable $\sup_{t \ge 0} B_t$. The idea is the following: let $a > 0$ and let T denote the first hitting time of a by the Brownian motion, where the process $B_t = B_{t-1} + B_t$ is a Brownian motion. We have, for any $t > 0$, $P(T < t) = P(T < t) = 1$. This implies, should still hold if we replace t by a stopping time independent of the process $(B_t)_{t \ge 0}$, for example $t = T$, and we should have $P(\sup_{t \ge 0} B_t \ge a) = P(T \le t) = 2P(T \le t)$. This is the first application of the Strong Markov Property. We now give a detailed proof.

9.5.10 *Theorem.* Let $(B_t)_{t \ge 0}$ be a Brownian motion and $a > 0$. Then

$$P(\sup_{t \ge 0} B_t \ge a) = 2P(B_t \ge a)$$

PROOF Let T be the first hitting time of a and define the stopping times T_n as in (9.5.5). Since the time T_n is independent of $B_{T_n+s} - B_{T_n}$ we have

$$\begin{aligned} P(M_t = a, B_t \le b) &= \sum_{n=0}^{\infty} P\left(M_{T_n} = a, B_{T_n} - a \le b - a, t - \frac{1}{n} < T_n \le t\right) \\ &= \sum_{n=0}^{\infty} P(M_{T_n} = a) P\left(B_{T_n} - a \le b - a, t - \frac{1}{n} < T_n \le t\right) \\ &= \frac{1}{2} P(T_a \le t) \end{aligned}$$

Since we have the approximation $P(B_t = b, T_a \le t) = P(B_t - a \le b - a, T_a \le t)$ and $P(T_a = t) = P(B_t = a) = 0$, an application of the reflection principle shows that

$$P(M_t \ge a, B_t \le b) = P(B_t - 2a \le b, T_a \le t) = \frac{1}{2} P(\sup_{0 \le s \le t} B_s \ge a, B_t \le b). \quad \square$$

Problems

1 Let $(B_t)_{t \ge 0}$ be a Brownian motion, $a > 0$, $A \in \mathcal{A}(B_s; s \le t)$ and $\Phi_{t-s}(B_s - x)$ as in (9.5.1). Show that

$$P(M_t \ge a, B_t \in A \mid B_t = y) = \int_{-\infty}^{2a-y} \mathbf{1}_A(z) p_{t-s}(y, z)\, dz \Big/ p_t(0, y),$$

where $p_t(x, y) = (1/\sqrt{2\pi t}) \exp(-(x-y)^2/2t)$.

(Hint: use the conditional density of $(B_s \mid B_t = y)$ from (9.5.1).)

2 In this problem we will show that Φ_t and $\mathcal{A}_{t+}$ differ only by sets of probability 0.

(a) If f is a measurable, bounded function from $\mathbb{R}$ into $\mathbb{R}$ and $s > 0$,

$$\varphi(s, x) = \int_{-\infty}^{\infty} f(y) \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{(y-x)^2}{2s}\right) dy.$$

Show that the function $(s, x) \to \varphi(s, x)$ is continuous.

(b) The Reflection Principle shows that if $s < t$ and f is a measurable, bounded function from $\mathbb{R}$ into $\mathbb{R}$, then

$$E^x(f(B_s); T_a \le s) = E^x(p_{t-s}(B_{T_a}, B_s) f(B_s); T_a \le s).$$

(c) Show that if $s < t$ and f is a measurable, bounded function from $\mathbb{R}$ into $\mathbb{R}$, then

$$E^x(f(B_s); T_a \le s) = E^x(p_{t-s}(B_{T_a}, B_s) f(B_s); T_a \le s).$$

(c) Show that if the functions $f_1, f_2, \ldots, f_n$ are measurable and bounded and if $1 < p < q < \infty$, then

$$E\left[\prod_{i=1}^{n} |f_i(X_i)|^p \,\middle|\, \mathcal{G}\right]^{1/p} \le E\left[\prod_{i=1}^{n} |f_i(X_i)|^q \,\middle|\, \mathcal{G}\right]^{1/q} \quad \text{a.e.}$$

(d) Show that the inequality in (c) is false when $p = 1$, even if restricted to being larger than 1.

(e) Now show that if the random variable Z is bounded and $\sigma(Z, \mathcal{G})$-measurable, then

$$E[ZY \mid \mathcal{G}] = E[ZY \mid \mathcal{H}],$$

where Y is any $\mathcal{H}$-measurable random variable for which ZY is integrable.

(f) Let $\mathcal{H}$ be the smallest complete σ-field generated by the random variables $X_1, \ldots, X_n$. Define $Y = \sum_{i=1}^{n} a_i 1_{(X_i = x_i)}$ where $P(X_i = x_i) > 0$, and let $\mathcal{H}_i$ be the σ-field generated by X_i, and show that the variable $Z = \prod_{i=1}^{n} 1_{(X_i = x_i)}$ is $\sigma(X_{n+1})$-measurable.

(Hint: Use Exercise 9.4.4 for parts (e) and (f).)

5. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\mathcal{F}_t)_{t \ge 0}$ be a nondecreasing family of sub-σ-fields of $\mathcal{A}$. Let $\mathcal{F}_\infty = \mathcal{F}_{[0,\infty)}$.

(a) Show that a nonnegative random variable T is a stopping time for the family $(\mathcal{F}_t)_{t \ge 0}$ if $\{T < t\} \in \mathcal{F}_t$ for any $t \ge 0$.

(b) Show that if T is a stopping time for the family $(\mathcal{F}_t)_{t \ge 0}$, we have $\mathcal{F}_{T-} = \mathcal{F}_T$.

6. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(\mathcal{F}_t)_{t \ge 0}$ be a nondecreasing family of sub-σ-fields of $\mathcal{A}$.

(a) Show that if T and S are two stopping times for the family $(\mathcal{F}_t)_{t \ge 0}$, then so are $[S < T]$, $[S \le T]$, $[S = T]$, $S \wedge T$, and $[S \ge T]$ are all in $\mathcal{F}_S \cap \mathcal{F}_T$.

(b) Show that, if A is a set in $\mathcal{F}_S$, then the set $A \cap [S \le T]$ is in $\mathcal{F}_T$.

(c) Show that if the set A is in $\mathcal{F}_S$, then the set $A \cap [S < T]$ is in $\mathcal{F}_T$.

(Hint: Use (a) to establish that $[S < T]$ and $[S \le T]$ are in $\mathcal{F}_T$. Use arguments similar to those of Theorem 9.5.3(b).)

The above results generalize properties such as "if S is a stopping time, the sets $[S < t]$, $[S = t]$, $[S \le t]$ are all in $\mathcal{F}_t$" and "if $A \in \mathcal{F}_t$, then $A \cap [S = t] \in \mathcal{F}_t$" in the case of a nonnegative stopping time T.

7. Let $(\Omega, \mathcal{A}, P)$ be a probability space, $(\mathcal{F}_t)_{t \ge 0}$ be a nondecreasing family of sub-σ-fields of $\mathcal{A}$, and (T_n) be a sequence of stopping times.

(a) Show that $\sup_n T_n$ is a stopping time.

...: Since this, in our family $(\mathcal{F}_t)_{t\geq 0}$ is right continuous, $(\mathcal{F}_t)$ is a stopping time and $\mathcal{F}_\tau = \bigcap_{t>0} \mathcal{F}_{\tau+t}$. Since τ is finite, $\mathcal{F}_\tau = \mathcal{F}_{\tau+}$.

9.6 Martingales

In Chapter 8, we defined the notion of a martingale in discrete time. We now consider processes $(X_t)_{t\geq 0}$ of random variables; in the second we generalize these notions to the case of an uncountable index set T. Many results of Chapter 8 can be generalized in a straightforward manner, and we will give only the results in Section 9.1. However, some aspects of the general theory of stochastic integrals will require the Brownian motion, B_t and $\mathcal{F}_t^B$.

9.6.1 Definition Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $(\mathcal{F}_t)_{t\geq 0}$ be a nondecreasing family of sub-σ-fields of $\mathcal{F}$. A family of random variables $(X_t)_{t\geq 0}$ is a martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$ if

(a) each X_t is $\mathcal{F}_t$-measurable and integrable, and
(b) $E[X_t \mid \mathcal{F}_s] = X_s$ a.s. if $0 \leq s \leq t < \infty$.

The notions of sub- and supermartingales can be similarly generalized.

If $\mathcal{F}_t = \sigma(X_s : s \leq t)$, we will say that $(X_t)_{t\geq 0}$ is a martingale without specifying the family of σ-fields.

9.6.2 Examples 1. Let Brownian motion $(B_t)_{t\geq 0}$ be a martingale. It is easy to check since the random variable $B_t - B_s$ is independent of $\mathcal{F}_s^B$ and has mean 0 if

$$0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n \leq s,$$

$$E[(B_{t_1} - B_{s_1})(B_{t_2} - B_{s_2}) \cdots (B_{t_n} - B_{s_n})] = 0.$$

2. The process $Y_t = B_t^2 - t$ is also a martingale as seen by observing that, if $s < t$,

$$E[B_t^2 - B_s^2 \mid \mathcal{F}_s^B] = E[(B_t - B_s)^2] = t - s,$$

and

$$E[(B_t - B_s)^2 \mid \mathcal{F}_s^B] = E[B_t^2] - 2E[B_t B_s] + B_s^2 \mid \mathcal{F}_s^B] = E[B_t^2 - B_s^2 \mid \mathcal{F}_s^B],$$

so that

$$E[B_t^2 - B_s^2 \mid \mathcal{F}_s^B] = t - s.$$

To work with the uncountable index set $\mathbb{R}^+$, we need some regularity conditions on martingales. This is why our first step is to show that martingales have continuous modifications, and our first theorem shows the existence of a right-continuous version of a martingale. This argument is similar to the proof of the existence of a continuous version of the Brownian motion. We first study the martingales on a countable dense subset S of $\mathbb{R}^+$ and we then construct a right-continuous version by taking limits along S.

9.6.3 **Theorem.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\{\mathcal{F}_t\}_{t\geq 0}$ a right-continuous non-decreasing family of sub-σ-fields. We assume that the probability space is complete in the sense of 1.7.7, and that each $\mathcal{F}_t$ contains the set $\mathcal{N} = \{A \in \mathcal{F} : P(A) = 0\}$. Then any martingale $\{X_t\}_{t\geq 0}$ with respect to $\{\mathcal{F}_t\}_{t\geq 0}$ admits a right-continuous version.

Proof. (a) Let S be a countable dense subset of $\mathbb{R}_+$, and b be a real number such that $a < b$, and let $S_n = S \cap [0, n]$. For any finite subset I of S_n, it follows from Theorem 9.6.1 that the number $U_I(a,b)$ of upcrossings of $[a,b]$ by the process $\{X_t\}_{t\in I}$ satisfies

$$E[U_I(a,b)] \leq \frac{1}{b-a} E[(X_n - a)^+],$$

and therefore, for any $\lambda > 0$,

$$P[U_I(a,b) \geq \lambda] \leq \frac{1}{\lambda(b-a)} E[(X_n - a)^+]. \tag{1}$$

Let $I_1 \subset I_2 \subset \cdots$ be an increasing sequence of finite subsets of S_n such that $\bigcup_k I_k = S_n$. The set $\{\omega : U_{S_n}(a,b)(\omega) = \infty\}$ is measurable, has probability 0, and contains the set $A = \{\omega : \exists t \in [0,n] \text{ such that } \inf_{\varepsilon>0} \sup_{s,u\in(t-\varepsilon,t+\varepsilon)\cap S_n} |X_s(\omega) - X_u(\omega)| > 0\}$. If $\omega \notin A$, then, since the $U_I(a,b)$ increase with I, for any $\lambda > 0$ there exists k such that $P[U_{I_k}(a,b) \geq \lambda] > 0$. For λ large, this contradicts (1). Since $\mathcal{F}$ is complete, the probability $P(A)$ is measurable and $P(A) = 0$.

(b) Let $\Lambda = \{\omega : \forall t \in [0,n], \exists \text{ sequence } t_k \to t \text{ such that } \lim_{k\to\infty} X_{t_k}(\omega) \text{ exists}\}$.

The set Λ is contained in the set $\bigcap_{a<b,\, a,b\in\mathbb{Q}} A_{a,b}$ where the union is over all integer values of a and all rational numbers $b > a$. Since $\bigcup_{a<b,\, a,b\in\mathbb{Q}} A_{a,b}$ has measure zero and has probability 0, Λ does also. Define the process

$$Y_t = \begin{cases} \lim_{s\to t,\, s\in S_n} X_s & \text{on the complement of } \Lambda, \\ 0 & \text{on } \Lambda. \end{cases}$$

Since $\mathcal{F}_t = \mathcal{F}_{t+}$ contains all events in $\mathcal{N} = \{A : P(A) = 0\}$, the random variable Y_t is $\mathcal{F}_t$-measurable.

In the equality $X_s = E[X_t \mid \mathcal{F}_s]$, let s decrease to t. $\mathcal{F}_s$ decreases as $s \to t$, and $E[X_t \mid \mathcal{F}_s]$ converges to $E[X_t \mid \bigcap_{s>t} \mathcal{F}_s] = E[X_t \mid \mathcal{F}_t] = X_t$ a.s. (9.4.4). Therefore $\{Y_t\}_{t\geq 0}$ is a right-continuous version of the martingale $\{X_t\}_{t\geq 0}$, and is itself a martingale. $\square$

We now give a couple of results necessary to prove 9.6.5, which will be used in the construction of stochastic integrals.

9.6.4 *Lemma.* Let $(X_t)_{t\geq 0}$ be a right-continuous supermartingale with respect to a family $(\mathcal{F}_t)_{t\geq 0}$ of subsets such that for any t, $X_t \in \mathcal{F}_t$. Then for any $\lambda > 0$ and $a < b$,

$$\lambda P\{\sup_{a\leq t\leq b} X_t \geq \lambda\} \leq \int_{\{\sup_{a\leq t\leq b} X_t \geq \lambda\}} X_a \, dP.$$

PROOF. We first prove the result when the index set is finite. We consider the supermartingale only for values of t in the set $\{a_1, a_2, \ldots, a_n\}$. Let T be the stopping time defined by

$$T = \begin{cases} \min\{a_i : X_{a_i} \geq \lambda\}, & \text{on the set } \{\max_i X_{a_i} \geq \lambda\}, \\ a_n, & \text{on the set } \{\max_i X_{a_i} < \lambda\}. \end{cases}$$

We have

$$\begin{aligned} \lambda P\{\max_i X_{a_i} \geq \lambda\} &\leq \sum_{i=1}^{n} \lambda P\{X_{a_i} \geq \lambda, T = a_i\} \leq \sum_{i=1}^{n} \int_{\{T=a_i\}} X_{a_i} \, dP \\ &\leq \sum_{i=1}^{n-1} \int_{\{T=a_i\}} X_{a_{i+1}} \, dP + \int_{\{T=a_n\}} X_{a_n} \, dP \\ &= \int_{\{\max_i X_{a_i} \geq \lambda\}} X_{a_n} \, dP. \end{aligned}$$

Choose an increasing sequence of numbers t_n such that $\bigcup_n t_n = \mathbb{Q}_+ \cap [a,b]$, where $\mathbb{Q}_+ = \{k2^{-l} : k, l \text{ are integers}\}$ and $t_1 = a$, $t_2 = b$. Since the random variable X_t is right continuous, and the sets $\{\max_{i\leq n} X_{t_i} \geq \lambda\}$ increase to the set $\{\sup_{t\in\mathbb{Q}_+} X_t \geq \lambda\}$, we obtain

$$\lambda P\{\sup_{t\in\mathbb{Q}_+} X_t \geq \lambda\} \leq \int_{\{\sup_{t\in\mathbb{Q}_+} X_t \geq \lambda\}} X_a \, dP. \tag{1}$$

Since the supermartingale $(X_t)_{t\geq 0}$ is right continuous, we have

$$\sup_{t\in\mathbb{Q}_+} X_t = \sup_{t\in[a,b]} X_t,$$

and

$$\lambda P\{\sup_{t\in[a,b]} X_t \geq \lambda\} \leq \int_{\{\sup_{t\in[a,b]} X_t \geq \lambda\}} X_a \, dP. \quad \square$$

9.6.5 Remark. If $(X_t)_{t\geq 0}$ is a right-continuous martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$, and each X^t is in L^2, then we have

$$\|\sup_{s\leq t} X_s\|_2 \leq 2\|X_t\|_2.$$

Proof. Let $Y = \sup_{s\leq t}|X_s|$. The process X_t is a submartingale, and therefore, by Doob's inequality, we have

$$\mathbb{E}[Y^2] \leq 2\mathbb{E}\left[\int_{[0,t]} X_s\, dY\right].$$

Since we do not know yet that $\|Y\|_2 < \infty$, we have to work with the random variables $Y_n = Y \wedge n$. The set $\{Y_n > s\}$ is empty if $s \geq n$, and otherwise coincides with $\{Y > s\}$ if $s < n$. Therefore Y_n satisfies

$$\mathbb{E}[Y_n^2] \leq 2\int_{[0,n)} |X_s|\, dY_s. \tag{1}$$

Let λ_n be the distribution function of Y_n. Since $Y_n = \int_0^n \mathbb{1}_{\{s<Y_n\}}\, dY_s$, it follows that

$$\begin{aligned}
\mathbb{E}[Y_n^2] &= \int_0^\infty s^2\, d\lambda_n(s) = \int_0^\infty \int_0^n 2s\, dY_s\, d\lambda_n(s) \\
&= \int_0^n 2s\,\mathbb{P}[Y_n \geq s]\, ds \qquad \text{by Fubini's Theorem} \\
&\leq 2\int_0^n \int_{[0,s]} |X_t|\, dY_t\, ds \qquad \text{by (1)} \\
&= 2\int_0^n |X_t| \int_t^n \mathbb{1}_{\{t\leq s\}}\, ds\, dY_t \\
&= 2\mathbb{E}[|X_t| Y_n] \leq 2\|X_t\|_2 \|Y_n\|_2,
\end{aligned}$$

and

$$\|Y_n\|_2 \leq 2\|X_t\|_2.$$

We finish by letting n approach ∞. $\square$

We now assume that the hypotheses of 9.5.3 are satisfied: $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and $(\mathcal{F}_t)_{t\geq 0}$ a non-decreasing right-continuous family of sub-σ-fields. The probability space is complete in the sense of 3.5.7, and each $\mathcal{F}_t$ contains the set $N = \{A \in \mathcal{F} : \mathbb{P}(A) = 0\}$.

9.6.6 *Definition.* We denote by $\mathcal{M}_T$ the vector space of all processes $X = (X_t)_{0 \le t \le T}$ such that $(X_t)_{0 \le t \le T}$ is a square-integrable, right-continuous martingale and X_T is in L^2.

Two elements X and Y of $\mathcal{M}_T$ are considered to be the same if $X_t(\omega) = Y_t(\omega)$ for almost every $(t, \omega) \in [0,T] \times \Omega$ (i.e. the set $\{(t,\omega) : X_t(\omega) \neq Y_t(\omega)\}$ is measurable and has probability 0 (see 9.1.7). In this case we will not distinguish between the processes X and Y.

We consider on $\mathcal{M}_T$ the inner product $\langle X, Y \rangle = E(X_T Y_T)$.

9.6.7 *Theorem.*

(a) If $\langle X, X \rangle = 0$, the martingale X is indistinguishable from the process identically equal to 0. Therefore $\| X \| = \langle X, X \rangle^{1/2}$ defines a norm on $\mathcal{M}_T$.

(b) If a sequence $(X^n)_{n \ge 1}$ converges to X in $\mathcal{M}_T$, then $\sup_{0 \le t \le T} |X^n_t - X_t|$ converges to 0 in L^2. In particular, for any $0 \le t \le T$, $X^n_t \to X_t$ in L^2.

Proof. This follows from Lemma 9.6.5.

9.6.8 *Theorem.* The space $\mathcal{M}_T$ is a Hilbert space for the inner product $\langle X, Y \rangle = E(X_T Y_T)$, and the subspace of continuous martingales is closed in $\mathcal{M}_T$.

Proof. Let (X^n) be a Cauchy sequence in $\mathcal{M}_T$. The sequence of terminal variables X^n_T is a Cauchy sequence in L^2; let Z be its L^2 limit, and $Z_t = E(Z \mid \mathcal{F}_t)$, $0 \le t \le T$. The martingale $(Z_t)_{0 \le t \le T}$ admits a right-continuous version $(Y_t)_{0 \le t \le T}$ (see 9.1.7), and the X^n_t converge to Y_t in L^2.

Assume now that the martingales X^n are continuous, and let X be their right-continuous limit in $\mathcal{M}_T$. We can assume that $\sum_n E(|X^n_T - X_T|^2)^{1/2} < \infty$ (if necessary, take a subsequence). Applying successively 9.6.5 and Chebyshev's inequality, we have

$$\sum_{n=1}^{\infty} P\left(\sup_{0 \le t \le T} |X^n_t - X_t| \ge \varepsilon\right) < \infty$$

and

$$\sum_{n=1}^{\infty} P\left(\sup_{0 \le t \le T} |X^{n+1}_t - X^n_t| \ge \varepsilon\right) < \infty$$

for any $\varepsilon > 0$. By the first Borel-Cantelli lemma we get

$$P\left(\limsup_{n \to \infty} \sup_{0 \le t \le T} |X^n_t - X_t| \ge \varepsilon\right) = 0.$$

and, taking a sequence $\varepsilon_k \downarrow 0$, we see that

$$P\{\sup_{t\le a}|X_t - X_s| \ge \varepsilon\} \le \sum_k P\{\text{[illegible]}\sup_{s\le t\le s+\delta_k}|X_t - X_s| \ge \varepsilon_k\} < \infty.$$

Hence almost all paths are right-continuous and $t \mapsto X_t(\omega) - \gamma t$ is a uniform limit of the continuous paths $t \mapsto X_{t_k}(\omega) - \gamma t$, and the martingale X is indistinguishable on $[0,a]$ from a continuous one. $\square$

Problems

1. Let $(X_k)_{k\ge 1}$ be a sequence of independent Gaussians such that $X_k \in L^1$ for all $k \ge 1$. Show that $Y_n = X_n - E(X_n)$ is a martingale.

2. (a) Let $(B_t)_{t\ge 0}$ be a Brownian motion and $\mathcal{F}_t = \sigma(B_s, s \le t)$. Show that, for each $\alpha \in \mathbb{R}$, the process
$$Y_t^{(\alpha)} = \exp(\alpha B_t - \tfrac{1}{2}\alpha^2 t)$$
is a martingale for the family $(\mathcal{F}_t)_{t\ge 0}$.

   (b) Conversely, let $(X_t)_{t\ge 0}$ be a continuous stochastic process such that $X_0 = 0$ and for each $\alpha \in \mathbb{R}$ the process
$$Y_t^{(\alpha)} = \exp(\alpha X_t + \tfrac{1}{2}\alpha^2 t)$$
is a martingale with respect to the family $\mathcal{F}_t = \sigma(X_s, s \le t)$. Show that $(X_t)_{t\ge 0}$ is a Brownian motion.

3. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathcal{F}_t)_{t\ge 0}$ a nondecreasing right-continuous family of sub-σ-fields and $(\mathcal{F}_t)_{t\ge 0}$ right-continuous martingales.

   (a) Show that if T is a stopping time bounded by t_0, then X_T is $\mathcal{F}_T$-measurable and $X_T = E(X_{t_0}|\mathcal{F}_T)$ a.s.

   (b) Show that if T is a bounded stopping time, then the process $Y_t = X_{t\wedge T}$, $t \ge 0$, is a martingale.

   (Hint: use the stopping times $T_n = k/n$ if $(k-1)/n \le T < k/n$, $1 \le k \le n$.)

   (c) [illegible] Theorem 3 of §3.)

4. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\mathcal{F}_t)_{t\ge 0}$ a nondecreasing right-continuous family of sub-σ-fields and $(X_t)_{t\ge 0}$ a right-continuous martingale. We recall that a random variable $X_\infty \in L^1(\Omega, \mathcal{F}, P)$ is uniformly integrable if

   (a) Show that $X_\infty = \lim_{t\to\infty} X_t$ exists a.s. and that $(X_t)_{t\ge 0}$ is a uniformly integrable martingale.

   (b) Show that for any stopping time T the random variable X_T is $\mathcal{F}_T$-measurable, integrable and $X_T = E(X_\infty|\mathcal{F}_T)$ a.s.

   (Hint: the assertions and methods of the preceding problem apply to right-continuous martingales.)

9.7 Itô integrals

One possible mathematical model for the price of shares on the stock market is the Brownian motion. The gain or loss in the value of a portfolio over a fixed time interval is assumed to equal the product of the number of shares held during the time interval and the change in the price of the shares over the time interval. At each time t_j the investor knows only what has happened up to time t_j. Therefore each random variable $f(t_j,\cdot)$ should be $\mathcal{F}_{t_j}$-measurable. What happens when the time intervals between consecutive decisions become smaller? Does the above sum have a limit? That is, can we define integrals of the form $\int_0^t f(s,\omega)\,dB_s(\omega)$? The paths of the Brownian motion have infinite variation on finite intervals, so we could use neither Lebesgue–Stieltjes integrals nor Riemann integrals. Indeed, almost all of the functions $s \mapsto B_s(\omega)$ would be of infinite variation on $[0,t]$, and hence integrals with respect to $dB_s(\omega)$ could be defined, for each ω, as Lebesgue–Stieltjes integrals. Unfortunately, this is not the case and we have to use a more complicated theory to get around the difficulty. The symbol $\int_0^t f\,dB_s$ for stochastic integrals will no longer be defined path by path, as a limit as $\Delta t \to 0$.

Let $(B_t)_{t\ge 0}$ be a Brownian motion on a probability space $(\Omega,\mathcal{F},P)$, and let $\mathcal{F}_t$ be the σ-algebra of events observable by the time t, generated by all the random variables B_s, $s \le t$. We will work with the σ-fields $\mathcal{N}$ of events of probability 0 and $\mathcal{F}_t^+ = \bigcap_{s>t} \mathcal{F}_s$. The family $(\mathcal{F}_t^+)_{t\ge 0}$ is right-continuous by Theorem 9.7, therefore the probability space $(\Omega,\mathcal{F},P)$ and the family $(\mathcal{F}_t^+)_{t\ge 0}$ satisfy the hypotheses of Theorems 9.6.3 and 9.6.6.

Which processes $f(t,\omega)$ do we want to integrate? First we want $\int_0^t f(s,\omega)\,dB_s$ to be a random variable, the process f should be measurable in both variables. We will also assume, as in the example of the investor, that for each t, the random variable $\omega \mapsto f(t,\omega)$ does not depend on the future. That is, we consider only processes $f(t,\omega)$ for which we can guess the following conditions:

(a) the process $f(t,\omega)$ is adapted, and
(b) the process $f(t,\omega)$ is measurable (jointly) for $\mathcal{B}\times\mathcal{F}$—as each t, the random variable $\omega \mapsto f(t,\omega)$ is $\mathcal{F}$-measurable.

For simple processes $f(t,\omega)$ we know what the integral should be.

(a) If $f(t,\omega) = I_{(a,b]}(t)$, then $\int_0^t f(s,\omega)\,dB_s = B_{t\wedge b} - B_{t\wedge a}$.
(b) If $f(t,\omega) = \sum_{j=1}^n I_{(t_{j-1},t_j]}(t)\,\xi_j(\omega)$, then $\int_0^t f(s,\omega)\,dB_s = \sum_{j=1}^n \xi_j(B_{t\wedge t_j} - B_{t\wedge t_{j-1}})$.

For the second we choose $a < b$ and we restrict ourselves to processes that vanish outside $[a,b]\times\Omega$.

9.7.2 *Definition*

(a) We denote by $\mathcal{H}_2$ the set of all processes $f(t,\omega) = f_1(t)\phi_1(\omega) + \cdots + f_n(t)\phi_n(\omega)$ where $f_1, \ldots, f_n$ are step functions on $[0,T]$ and

(i) $0 = t_1 < t_2 < \cdots < t_{n+1} = T$ is a finite partition of $[0,T]$,

(ii) ϕ_i is $\mathcal{M}^i$ measurable,

(iii) $E\phi_i^2 < \infty$, $i = 1, \ldots, n$, and

(iv) each f_i is an L^2 function on $[t_i, t_{i+1})$.

For such a process we define, for $0 \le t \le T$,

$$I_t = \int_0^t f(s,\omega)\,dB_s(\omega) = \sum_{i=1}^{n} \phi_i(\omega)[B_{t_{i+1}}(\omega) - B_{t_i}(\omega)].$$

(b) We denote by $\mathcal{H}_2$ the set of measurable martingale type processes $g(t,\omega)$ for which there exists a sequence $\{g_n\}$ in $\mathcal{H}_2$ such that $E\int_0^T [g_n(t,\omega) - g(t,\omega)]^2\,dt \to 0$.

For processes in $\mathcal{H}_2$ we have the following straightforward lemma.

9.7.2 *Lemma.* If f and g are two processes in $\mathcal{H}_2$ then the processes $I_t = \int_0^t f(s,\omega)\,dB_s$ and $J_t = \int_0^t g(s,\omega)\,dB_s$, $0 \le t \le T$, satisfy the following properties:

(a) $\{I_t, \mathcal{M}_t, 0 \le t \le T\}$ and $\{I_t^2 - \int_0^t f^2(s,\omega)\,ds, \mathcal{M}_t, 0 \le t \le T\}$ are continuous martingales for the family of σ-fields $\{\mathcal{M}_t\}_{0 \le t \le T}$.

(b) The martingales I and J are L^2-bounded, and for $0 \le s \le t \le T$ we have $E[I_t J_t] - E[I_s J_s] = E\left[\int_s^t f(u,\omega)g(u,\omega)\,du\right]$.

(c) In particular $E[I_T^2] = E\left[\int_0^T f^2(u,\omega)\,du\right]$, which we write as $\|I\|_2 = \|f\|_{L^2}$.

(d) $E\left[\sup_{0 \le t \le T} |I_t - J_t|^2\right] \le 4E\left[\int_0^T |f(u,\omega) - g(u,\omega)|^2\,du\right]$.

Proof. If the ϕ_i's are in L^2 and ϕ_i is $\mathcal{M}_s$ measurable, then we have

$$E[\phi_i(B_{t_{i+1}} - B_{t_i}) \mid \mathcal{M}_s] = \phi_i E[B_{t_{i+1}} - B_{t_i} \mid \mathcal{M}_s] = 0$$

$$E[\phi_i^2(B_{t_{i+1}} - B_{t_i})^2 \mid \mathcal{M}_s] = E[\phi_i^2 E[(B_{t_{i+1}} - B_{t_i})^2 \mid \mathcal{M}_{t_i}] \mid \mathcal{M}_s] = E[\phi_i^2(t_{i+1} - t_i) \mid \mathcal{M}_s]$$

and for $i < j$

$$E[\phi_i\phi_j(B_{t_{i+1}} - B_{t_i})(B_{t_{j+1}} - B_{t_j})$$

$$= E[\phi_i\phi_j(B_{t_{i+1}} - B_{t_i})E[B_{t_{j+1}} - B_{t_j} \mid \mathcal{M}_{t_j}]] = 0.$$

9.7.3 Theorem.

(a) The mapping $\chi \to U\chi = \int_0^t \chi(s)dW(s)$, $\chi \in \mathcal{H}_2^0$, is an isometry from $\mathcal{H}_2^0$ onto $\mathcal{H}_2^0$ which commutes with the action of the Hilbert space $L^2(\Omega, \mathcal{F}, P)$ on $\mathcal{H}_2^0$ and $\mathcal{H}_2^0$ (i.e. $U(\xi\chi) = \xi U\chi$ for all $\xi \in L^2(\Omega, \mathcal{F}, P)$ and $\chi \in \mathcal{H}_2^0$). The mapping can therefore be extended to the closure in $\mathcal{H}_2$.

(b) The elements of $\mathcal{H}_2^0$ on $\mathbb{R}^1[0,\infty) \times [0,\infty) \times \Omega$ are $\mathcal{P}$-measurable and $\mathcal{H}_2$ contains $\mathcal{H}_2^0$.

(c) If $\chi \in \mathcal{H}_2$ for a process $Y_t = \int_0^t \chi(s)dW_s$ thus defined is a P-unit in $\mathcal{M}_2$, is a continuous martingale, and, for each $t > 0$, $E[Y_t^2] = E[\int_0^t \chi^2(s)ds]$.

(d) If χ and η are two processes in $\mathcal{M}_{2,0}$ for which $Y_t = \int_0^t \chi(s)dW_s$ and $Z_t = \int_0^t \eta(s)dW_s$, $0 \le t < \infty$, satisfy $E[Y_t^2] = E[\int_0^t \chi^2(s)ds]$,

(e) If $\chi \in \mathcal{H}_2$ and τ is a stopping time, $\int_0^\tau \chi(s)dW_s = \int_0^\infty \chi(s)\mathbb{1}_{[0,\tau]}(s)dW_s$ a.s.

Part (a) is a consequence of (c) and (d), and (e) is a consequence of (b) and (c). The proofs of (b) and (c) should be clear from the definitions, so we just have to prove (d). We will follow the proof in Doob (1953).

1. Let $\chi(s)$ be a bounded Lebesgue measurable function which vanishes outside a finite interval. Then, by 2.4.5, for every $\varepsilon > 0$ there exists a continuous, bounded function η_ε which vanishes outside a finite interval and such that

$$\int_{-\infty}^{\infty} |\chi(s) - \eta_\varepsilon(s)|^2 ds \le \varepsilon^2.$$

[By 2.4.5 there exists a continuous function $\tilde{\eta}_\varepsilon$ such that $\int_{-\infty}^{\infty} |\chi(s) - \tilde{\eta}_\varepsilon(s)|^2 ds < \varepsilon^2$. Assume that the function $\tilde{\eta}_\varepsilon$ vanishes outside a finite interval $[a, b]$. Let η_ε be the continuous function which is 1 on $[a, b]$, 0 outside of $[a - \varepsilon, b + \varepsilon]$, and linear elsewhere. The function $\eta_\varepsilon \tilde{\eta}_\varepsilon$ vanishes outside a finite interval and $\int_{-\infty}^{\infty} |\chi(s) - \eta_\varepsilon(s)\tilde{\eta}_\varepsilon(s)|^2 ds \le 2\int_{-\infty}^{\infty} |\chi(s) - \tilde{\eta}_\varepsilon(s)|^2 ds < 2\varepsilon^2$.]

Using Minkowski's inequality and the dominated convergence theorem we obtain

$$\limsup_{\varepsilon \to 0} \left| \int_{-\infty}^{\infty} \chi(s)\eta_\varepsilon(s)dW_s \right|^{1/2}$$

$$\le \limsup_{\varepsilon \to 0} \int_{-\infty}^{\infty} |\chi(s) - \eta_\varepsilon(s)|^2 ds \Big|^{1/2} \sup_{0 \le t < \infty} |W_t| = 0$$

so that

$$\lim_{\varepsilon \to 0} \int_{-\infty}^{\infty} \chi(s)\eta_\varepsilon(s)dW_s = U\chi.$$

7. Assume now that $\{g(t,x)\}$ is a jointly measurable, square integrable process that vanishes for $t \notin [0,a]$. We want to show that g is in the closure of $\mathcal{E}$.

We partition $\mathbb{R}$ into intervals $[x^{(n)}_j, x^{(n)}_{j+1}]$ and consider the process

$$g_n(t,x) = g(\alpha_n(t), x) \quad \text{where } \alpha_n(t) = t^{(n)}_j \text{ on } [t^{(n)}_j, t^{(n)}_{j+1}).$$

For fixed ω and x the process $g_n(t,x)$ is equal to $g(t^{(n)}_j, x)$ on the interval $[t^{(n)}_j, t^{(n)}_{j+1}]$. Therefore the restriction of $g_n(t,x,\omega)$ to $[0,a] \times \mathbb{R}$ is measurable with respect to $\mathcal{P}$. We now show that g_n is the L^2 limit of a sequence in $\mathcal{E}$. We have

$$\begin{aligned} &\lim_{k\to\infty} \left| \int_0^a \int_{-k}^{k} |g_n(t,x,\omega) - g(t,x,\omega)|^2 \, dx \, dt \right| \\ &\qquad = \lim_{k\to\infty} E \int_{-k}^{k} \int_0^a |g_n(t,x,\omega) - g(t,x,\omega)|^2 \, dt \, dx = 0 \end{aligned}$$

by the bounded convergence theorem. This will show that the process g is in the closure of $\mathcal{E}$.

Now it suffices to show that for each ω

$$\lim_{h\to 0} \int_{-\infty}^{\infty} |g(x+h,\omega) - g(x,\omega)|^2 \, dx = 0.$$

Since Lebesgue measure is translation invariant we have for each fixed t and ω

$$\lim_{h\to 0} \int_{-\infty}^{\infty} |g(t, x-h, \omega) - g(t,x,\omega)|^2 \, dx = 0.$$

We can replace ω by $\alpha_s(\omega) = \omega$, which approximates 0 as $s \to 0$, and obtain

$$\lim_{h\to 0} \int_0^a \int_{-\infty}^{\infty} |g(s, t, x-h, \omega) - g(s,t,x,\omega)|^2 \, dx \, dt = 0.$$

Since all the processes involved are bounded and vanish outside a fixed finite interval, we have

$$\begin{aligned} &\lim_{h\to 0} \int_{-\infty}^{\infty} \int_0^a \int_{-\infty}^{\infty} |g(s,t,x-h,\omega) - g(s,t,x,\omega)|^2 \, dx \, dt \, ds \\ &\qquad = \lim_{h\to 0} E \int_0^a \int_0^a \int_{-\infty}^{\infty} |g(s,t,x-h,\omega) - g(s,t,x,\omega)|^2 \, dx \, dt \, ds = 0. \end{aligned}$$

Therefore, we can choose a subsequence n_k such that

$$\lim_{h\to 0} E \left[\int_0^a |g(s, t+h, \omega) - g(s,t,\omega)|^2 \, dt \right] = 0 \quad \text{a.s. on } \Omega \times [0,a].$$

We have now shown that there exists a subsequence η_j and, for each x, a set Ω_x such that

$$\lim_{j\to\infty} E\left[\int_0^\infty |\sigma(\eta_j(t)-x)+\eta_j(t)-g(x,\omega)|^2\,dt\right]$$
$$= \lim_{j\to\infty} E\int_0^\infty \chi_{\Omega_x}(\omega)\,|\eta_j(t)-x|^2\,dt = 0$$

and η_j is in $\mathcal{H}^2$ for each j.

3. Take the process g on $[0,\infty)$ to be any $\{\mathcal{F}_t\}$-adapted, measurable process satisfying

$$E\left[\int_0^\infty g^2(t)\,dt\right] < \infty,$$

and define the processes

$$\eta_n = \begin{cases} g & \text{if } |g| \le n, \\ 0, & \text{otherwise.} \end{cases}$$

According to Part 2, each η_n is in the closure of $\mathcal{H}_0^2$. This dominated convergence theorem ensures that

$$E\left[\int_0^\infty |g(t) - \eta_n(t)|^2\,dt\right] \to 0$$

and gives the conclusion of Part 1.

The stochastic integrals $I_t = \int_0^t g(s)\,dB_s$ that we have thus defined as limits in $\mathcal{H}^2$ are random variables. They are only determined a.s. (Lemma 9.2.3 assures the existence of continuous versions of $\{I_t\}_{t\ge0}$, and when we talk about stochastic integrals, we always assume that we are working with a continuous version. It does not matter which, since two continuous versions are indistinguishable.

We now generalize the definition of stochastic integrals to the interval $[0,\infty)$.

9.2.4 Theorem. *For each $t\ge0$ let $\mathcal{F}_t$ be the set of measurable, real-valued stochastic processes $g(t,\omega)$ on $[0,\infty)\times\Omega$ such that $E\int_0^t g^2(s)\,ds < \infty$, for any $\omega > 0$.*

(a) There exists a unique measurable $\{\mathcal{F}_t\}_{t\ge0}$-adapted process $V_t = \int_0^t g(s)\,dB_s$ a.s. for any $t\ge0$.

(b) If f and g are in $\mathcal{H}$, then $\int_0^t (af+bg)\,dB_s = a\int_0^t f(s)\,dB_s + b\int_0^t g(s)\,dB_s$ a.s. for any t and any constants a and b.

(a) If $f \in \mathcal{L}^2$, then $\int_0^t f(s)\,dB_s = \int_0^t f(s)\,dW_{\varphi(s)}$ a.s. for all $t \geq 0$.
(c) If τ is any stopping time with $\mathbb{E}[\tau] < \infty$, then for any $t \geq 0$

$$\mathbb{E}\left[\left(\int_0^{\tau} f(s)\,dB_s\right)^2\right] = \mathbb{E}\left[\left(\int_0^{\tau} g(s)\,dW_s\right)^2\right] = \mathbb{E}\left[\int_0^{\tau} g^2(s)\,ds\right]$$

Proof. Property (a) follows from (c) and Exercise 7.3.8 by taking $\tau = \tau_n = \inf\{t \geq 0 : \int_0^t f^2(s)\,ds \geq n\}$, using the results in 7.3.1. This is possible since $f \in \mathcal{L}^2[0,t]$ implies $\int_0^t f^2(s)\,ds < \infty$ a.s. Property (b) of Theorem 7.2.3 ensures that, for $|h| < t$, the processes $\{Y_{t-h}\}$ and $\{Y_{t+h}\}$ are indistinguishable, and we can define the stochastic integrals $Z_t = \int_0^t f(s)\,dB_s$ for any $t \geq 0$ by taking $Z = Z_{t+1}$ on $[0, t+1]$. The process Z is continuous on $[0,\infty)$ and coincides with Y except possibly on the set $\bigcup_{n=1}^{\infty} \{\omega : Y_{t+1}(\omega) \neq Y_{n+1}(\omega)\}$, which has probability 0. We can therefore change the values of the Y_t's on this set (for example take $Y_t \equiv 0$) to get a continuous version of the stochastic integrals.

The other properties are true for any elementary process and extend to $\mathcal{L}^2$ as in Theorem 7.2.3.

Problems

1. Let W be a stopping time for an increasing family of σ-fields $\{\mathcal{N}_t\}_{t \geq 0}$ and assume that W is finite, so that the stochastic interval $[0, W) = \{(t, \omega) : 0 \leq t < W(\omega)\}$.
   (a) Show that the left-continuous process $\{Y_t\}$ is measurable and non-anticipating. Therefore if $f \in \mathcal{L}^2$ we can consider the two integrals $X_t = \int_0^t f(s)\,dB_s$ and $X = \int_0^W f(s)\,dB_s$.
   (b) Show that if $\{\tau_n\}$ and $\{\sigma_n\}$ are two sequences of stopping times which both approximate W from below (i.e. $\tau_n < W$ and $\tau_n \to W$ a.s., and similarly for σ_n), then the sequences $\{X_{\tau_n}\}$ and $\{X_{\sigma_n}\}$ are indistinguishable. (Show it is enough to consider the case when $\tau_n \leq \sigma_n$ and then use only a finite number of terms.)
   (c) Extend this property to the processes $f \in \mathcal{M}_W$ and then to the process $f = A$.
2. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\{\mathcal{N}_t\}_{t \geq 0}$ a nondecreasing, right-continuous family of sub-σ-fields of $\mathcal{F}$. A process $\{A_t\}_{t \geq 0}$ is called *progressively measurable* with respect to $\{\mathcal{N}_t\}$ if the restriction of A to $[0,t] \times \Omega$ is $\mathcal{B}[0,t] \times \mathcal{N}_t$ measurable.
   (a) Show that a progressively measurable process is measurable and non-anticipating.
   (b) Show that right-continuous and non-anticipating processes are progressively measurable.
3. Let $\{Y(t,\omega)\}$ be a progressively measurable process and $W(\omega)$ a stopping time for each $\omega \in \Omega$, $\{Y^W(t,\omega) = Y(W(\omega), \omega)\}$ is not necessarily measurable. We want to extend the notion of stochastic integral to this process Y.

(a) For each integer $n \geq 0$ let $T_n(\omega) = \inf\{t : \int_0^t f^2(s,\omega)\,ds > n\}$ ($T_n = \infty$ if the set $\{t : \int_0^t f^2(s,\omega)\,ds > n\}$ is empty.) Show that the T_n are stopping times for the family of σ-fields $\mathcal{F}_t^{(p)}$ and that $\lim_n T_n = \infty$ a.s.

(b) We denote by $[Y,Z]$ the stochastic process $(\langle Y+Z\rangle - \langle Y-Z\rangle)/4$. Show that $\int_0^{t\wedge T_n} f(s)\,dB_s$ is in $\mathcal{M}^2$.

(c) Let $Z_{t\wedge T_n} = \int_0^{t\wedge T_n} f(s)\,dB_s$ as defined in Theorem 9.2.4. Show that the processes $\{Z_{t\wedge T_n}\}_t$ and $\{Z_{t\wedge T_n}^2 - \int_0^{t\wedge T_n} f^2(s)\,ds\}_t$ are martingales.

(d) Show that there exists a unique increasing process $\{F_t\}_t$ such that for each n the processes $\{Z_{t\wedge T_n}^2 - F_{t\wedge T_n}\}_t$ and $\{Z_{t\wedge T_n}^2 - \int_0^{t\wedge T_n} f^2(s)\,ds\}_t$ are submartingales.

We now define $\int_0^t f(s)\,dB_s$ by $\int_0^t f(s)\,dB_s = F_t$. The process $\{F_t\}_t$ is not necessarily a martingale, but it is a local martingale in the sense that there exists a nondecreasing sequence of stopping times $T_1 \leq T_2 \leq \cdots \leq T_n \leq \cdots$ such that $\lim_n T_n = \infty$ and each $\{F_{t\wedge T_n}\}_t$ is a martingale.

(e) Let T be a stopping time such that $\int_0^T f^2(s)\,ds \leq K$ for $T_n = \inf\{t : \int_0^t f^2(s)\,ds > n\}$ as defined in 9.2.4. Use Problem 3 to show that the processes $\{Z_{t\wedge T_n}\}_t$ and $\{Z_{t\wedge T_n}^2 - \int_0^{t\wedge T_n} f^2(s)\,ds\}_t$ are martingales.

(f) Let $S_n(\omega) = 0, 1, \ldots$ be a nondecreasing sequence of stopping times such that $\lim_n S_n = \infty$ and such that $\{Z_{t\wedge S_n}\}_t$ is in $\mathcal{M}^2$. (Just as in (d), we could use the stopping times S_n to define the integrals $\int_0^t f(s)\,dB_s$, show that the integrals do not depend on the sequence of stopping times used in the construction.

5. Generalize Problem 3 to obtain stochastic integrals with respect to martingales such that, for each $t \geq 0$, $\int_0^t f^2(s)\,ds < \infty$ a.s. (Use an argument similar to the proof of 9.2.5 to show that the integral $\int_0^t f^2(s)\,ds$ exists in the Riemann sense.)

9.3 Ito's Differentiation Formula

Assume that a function $g(x)$, defined on $\mathbf{R}$, can be expanded in a Taylor series. We know that $g(t) = \int_0^t g'(s)\,ds$. We can heuristically justify this statement as follows. Let $0 = t_0 < t_1 < \cdots < t_n = t$ be a partition of $[0,t]$. Then

$$g(t) - g(0) = \sum_{i=0}^{n-1} [g(t_{i+1}) - g(t_i)]$$

$$= \sum_{i=0}^{n-1} \left[g'(t_i)(t_{i+1} - t_i) + \sum_{k=2}^{\infty} \frac{g^{(k)}(t_i)}{k!} (t_{i+1} - t_i)^k \right].$$

The first term $\sum_i g'(t_i)(t_{i+1} - t_i)$ converges to $\int_0^t g'(s)\,ds$, and for $k \geq 2$ each term $\sum_i g^{(k)}(t_i)(t_{i+1} - t_i)^k$ should go to 0 since $\sum_i (t_{i+1} - t_i)^k \leq \max_i (t_{i+1} - t_i)^{k-1} \sum_i (t_{i+1} - t_i) \to 0$ when the partition of $[0,t]$ gets finer

Therefore, if everything goes well, the sum over i of these terms should go to 0. This heuristic reasoning would still apply if we considered a continuous function $G(t)$ having finite variation on any finite interval:

$$\int_0^t G(s)\,df(s) = \int_0^t G(s)f'(s)\,ds = \sum_i \left[G(t_{i+1})f(t_{i+1}) - f(t_i)G(t_i)\right]$$
$$= \sum_i \left[f(t_i)\left(G(t_{i+1}) - G(t_i)\right) + \sum_{r=1}^{n} \frac{f^{(r)}(t_i)}{r!}\left(t_{i+1} - t_i\right)^r G(t_{i+1})\right].$$

The last term converges to $\int_0^t f(s)\,dG(s)$, and for each $r \geq 2$, $\sum_i f^{(r)}(t_i)(t_{i+1} - t_i)^r G(t_{i+1})$ should go to 0 since $\sum_i G(t_{i+1}) - G(t_i)|^2 \leq \max_i |G(t_{i+1}) - G(t_i)| \cdot \sum_i |G(t_{i+1}) - G(t_i)|$, and $\sum_i |G(t_{i+1}) - G(t_i)|$ remains bounded when the partition becomes fine (G is continuous and has finite variation on $[0,t]$).

What happens when we replace $G(s)$ by the paths B_s of the Brownian motion? The first term $\sum_i f(t_i)(B_{t_{i+1}} - B_{t_i})$ surely converges to $\int_0^t f(s)\,dB(s)$. The term $\sum_i f'(t_i)f(t_{i+1} - t_i)^2$ no longer converges to 0, since the variation of the paths is so sufficient. The sum $\sum_i (B_{t_{i+1}} - B_{t_i})^2$ converges in L^2 to t, so $\sum_i f'(t_i)f(t_{i+1} - t_i)^2$ certainly converges to $\int_0^t f'(s)\,ds$. Just as $\sum_i f'(t_i)(B_{t_{i+1}} - B_{t_i})$ should go to 0 since $\sum_i |B_{t_{i+1}} - B_{t_i}| \leq \max_i |B_{t_{i+1}} - B_{t_i}| \cdot \sum_i |B_{t_{i+1}} - B_{t_i}|$, now that we have multiplied by a second factor of $B_{t_{i+1}} - B_{t_i}$ we can only say that the sum goes to 0.

9.8.7 *Ito's Differentiation Formula.* Let f be a continuous function on $\mathbb{R}$. We assume that its first and second derivatives f' and f'' exist and are continuous. Then, for any $t > 0$,

$$f(B_t) - f(B_0) = \int_0^t f'(B_s)\,dB_s + \frac{1}{2}\int_0^t f''(B_s)\,ds \quad \text{a.s.}$$

Since the processes $Y_t = f(B_t) - f(B_0)$ and $X_t = \int_0^t f'(B_s)\,dB_s + \frac{1}{2}\int_0^t f''(B_s)\,ds$ are continuous, they will be indistinguishable if

Proof. 1. The first step is to make sure that the integrals exist in the sense of Theorem 9.8.1 exist. The process $f'(B_t)\mathbb{1}_{[0,t]}(t)$ is jointly measurable and continuous, therefore it is progressively measurable (Problem 2 of 9.7). Furthermore, the function $f'(B_s(\omega))$ is continuous in s everywhere, on any interval $[0,t]$, it is bounded by a constant $C(\omega, t)$ and $\int_0^t f'(B_s)^2\,ds < \infty$ is finite. Consequently we can, by Problem 3 of 9.8, define a continuous version of the

stochastic integrals $\int_0^t f(B_s)\,dB_s$. A similar but easier argument shows that the Lebesgue–Stieltjes integrals $\int_0^t f'(B_s)\,ds$ exist.

2. Let $\alpha \geq 0$ and $t_n = \|\alpha_1, \alpha_2, \ldots, \alpha_n = t$ be a partition $\mathcal{P}$ of $[0,t]$. We denote by $|\mathcal{P}|$ the quantity $\max_i |\alpha_{i+1} - \alpha_i|$. Since f is C^2 and f' is continuous, there exists a value η_{i+1} in between B_{α_i} and $B_{\alpha_{i+1}}$ such that

$$f(B_{\alpha_{i+1}}) - f(B_{\alpha_i}) = f'(B_{\alpha_i})(B_{\alpha_{i+1}} - B_{\alpha_i}) + \tfrac{1}{2} f''(\eta_{i+1})(B_{\alpha_{i+1}} - B_{\alpha_i})^2.$$

Then notice $\eta_{i+1}(\omega) \in [B_{\alpha_i}(\omega), B_{\alpha_{i+1}}(\omega)]$ such that $\eta_{i+1}(\omega) \to B_{\alpha_i}(\omega)$ for [this is due to the convergence of the function $x \mapsto B_x(\omega)$; the random variable η_i is not necessarily a stopping time], and we have

$$\begin{aligned} f(B_t) - f(B_0) &= \sum_i [f(B_{\alpha_{i+1}}) - f(B_{\alpha_i})] \\ &= \sum_i f'(B_{\alpha_i})(B_{\alpha_{i+1}} - B_{\alpha_i}) + \sum_i \tfrac{1}{2} f''(\eta_{i+1})(B_{\alpha_{i+1}} - B_{\alpha_i})^2. \end{aligned}$$

We study separately the sum of each term as the partition gets finer.

3. We want to show that $Z_t = \sum_i f'(B_{\alpha_i})(B_{\alpha_{i+1}} - B_{\alpha_i})$ converges to $U_t = \int_0^t f'(B_s)\,dB_s$. The function $\sum_i f'(B_{\alpha_i})\mathbf{1}_{[\alpha_i,\alpha_{i+1}[}(s)$ converges a.s. to the process $f'(B_s)\mathbf{1}_{[0,t]}(s)$, but since $\sup_{\alpha_i \leq t} |f'(B_{\alpha_i})|$ is not necessarily finite, we cannot argue that

$$E\left[\int_0^t \left(\sum_i f'(B_{\alpha_i})\mathbf{1}_{[\alpha_i,\alpha_{i+1}[}(s) - f'(B_s)\mathbf{1}_{[0,t]}(s)\right)^2 ds\right] \to 0.$$

Construct the times T_n as in Lemma 1 defined as follows:

$$T_n = \begin{cases} \inf\{s \geq 0 : |B_s| \geq n\}, & \text{if } \sup_{s \leq t} |B_s| \geq n, \\ t, & \text{if } \sup_{s \leq t} |B_s| \geq n - 1. \end{cases}$$

Take also $T_0 = 0$, $T_\infty = \sup_{s \leq t} |B_s| < n$. By Lemma 1, $T_n \uparrow t$, and the T_n form a nondecreasing sequence of stopping times that converges to t. For each n, the function $f'(x)$ is bounded on $[-n, n]$. Therefore

$$E\left[\int_0^{T_n} \left(\sum_i f'(B_{\alpha_i})\mathbf{1}_{[\alpha_i,\alpha_{i+1}[}(s) - f'(B_s)\right)^2 ds\right] \to 0,$$

where $\|\mathcal{P}\| \to 0$; there exist $\mathcal{P}_n$ such that $\|\mathcal{P}_n\| \to 0$ in L^2 and in probability [see Lemma Problem 9.7]. The terms Z_t converge, and $U_t = U_{t \wedge T_n}$, the convergence

$\ldots$ and $\|\theta' - \theta\| < \delta$, then $P(\theta' T_{n-1} + (1-\theta')T_n \in C) \ge \alpha$. Since θ' is small enough to have $P(\theta' T_{n-1} + (1-\theta')T_n \in C) \ge \alpha$, we obtain

$$P(|\theta_n - \theta| < \varepsilon) \ge P(T_{n-1} - \varepsilon < \theta < T_n + \varepsilon) = P(T_{n-1} < \theta + \varepsilon, \theta - \varepsilon < T_n) \ge 2\alpha - 1.$$

Therefore θ_n converges in probability to θ.

A somewhat stronger theorem is given by Wald [1949]. We state it here to show that if the sequence of partitions is well chosen, consistency follows from $\int_\Omega f^*(\theta, x)\,dx$ in probability.

For partitions $\mathcal{P}$ of $[0,1]$, let us consider the random variables

$$\Delta(\mathcal{P}) = \sum_{\mathcal{P}} f^*(\theta_{i+1}) - f^*(\theta_i) \cdot (\theta_{i+1} - \theta_i)^2,$$

$$\gamma(\mathcal{P}) = \sum_{\mathcal{P}} f^*(\theta_i, \theta_{i+1}) \cdot (\theta_{i+1} - \theta_i)^2 \cdot \theta_i (1 - \theta_i).$$

Since $\sum_{\mathcal{P}} f^*(\theta_i, \theta_{i+1}) \cdot (\theta_{i+1} - \theta_i)^2$ converges for each ω in $\int_0^1 f^*(\theta, x)\,dx < \infty$, it is enough to show that $\sum_{\mathcal{P}} f^*(\theta_i, \theta_{i+1}) \cdot (\theta_{i+1} - \theta_i)^2 - \sum_{\mathcal{P}} f^*(\theta_i, \theta_{i+1}) \cdot (\theta_{i+1} - \theta_i)\theta_i(1-\theta_i) = \Delta(\mathcal{P}) - \gamma(\mathcal{P})$ converges to 0 in probability.

We first study the term $\Delta(\mathcal{P})$:

$$|\Delta(\mathcal{P})| < \max_{\mathcal{P}} |f^*(\theta_{i+1}) - f^*(\theta_i)| \cdot \sum_{\mathcal{P}} (\theta_{i+1} - \theta_i)^2.$$

According to 9.8.1, we can choose a sequence of partitions $\mathcal{P}_n$, $n \ge 1$, of $[0,1]$ such that $|\mathcal{P}_n| \to 0$ and $\sum_{\mathcal{P}_n} (\theta_{i+1} - \theta_i)^2$ converges, since $\int_0^1 f^*(\theta, x)\,dx$ exists. This implies the existence of a subsequence which converges to 0. For each ω, the continuous function $\theta \to f^*(\theta, x)$ is uniformly continuous on $[0,1]$, and $\max_{\mathcal{P}} |f^*(\theta_{i+1}) - f^*(\theta_i)| \to 0$ when $|\mathcal{P}_n| \to 0$. Therefore $\Delta(\mathcal{P}_n) \to 0$ a.s. and in probability.

We now study the term $\gamma(\mathcal{P})$. Let $T_{n-1} < \theta < T_n$, the last stopping time defined in part 3 of the proof. For each ω, we have

$$\sup_{\theta_i \le \theta} f^*(\theta_i, \theta_{i+1}) = \sup_{\theta_i \le \theta} f^*(\theta_i, X_n).$$

We construct the random variables

$$Z(\omega) = \sum_{\mathcal{P}} f^*(\theta_i, \theta_{i+1})(\theta_{i+1} - \theta_i)^2 \cdot \theta_i(1-\theta_i) \cdot 1_{\{\theta_{i+1} < T_n\}} \cdot 1_{\{\theta_i > T_{n-1}\}},$$

and denote by V_i the random variable

$$V_i = f^*(\theta_{i+1}, X_n)(\theta_{i+1} - \theta_i)^2 \cdot \theta_i(1-\theta_i) \cdot 1_{\{\theta_{i+1} < T_n\}} \cdot 1_{\{\theta_i > T_{n-1}\}}.$$

We have

$$\mathbb{E}[|Y_t^n|^2] = \mathbb{E}\left[\sum_{i=1}^{d}\left(\int_0^t \sigma_i^n(s)\,dB_s^i\right)^2\right] + \mathbb{E}\left[\sum_{i=1}^{d}\int_0^t |b_i^n(s)|^2\,ds\right] + 2\sum_{i<j}\mathbb{E}[A_t^{ij}A_t^{ji}],$$

since the processes $B_{t\wedge\tau_n}$ and $C_t = B_t^i B_t^j - \delta_{ij}t$ are martingales for the family of σ-fields $\mathcal{F}_{t\wedge\tau_n}$ (apply Problem 3 of 9.6 to the random variables B_{t_1}, B_{t_2} and the bounded stopping time $T = t\wedge\tau_n$), we have

$$\begin{aligned}
\mathbb{E}[B_{t_1\wedge\tau_n}B_{t_2\wedge\tau_n} \mid \mathcal{F}_{t_1\wedge\tau_n}] &= \mathbb{E}[B_{t_1\wedge\tau_n}^2 - 2B_{t_1\wedge\tau_n}B_{t_1\wedge\tau_n} + B_{t_1\wedge\tau_n}^2 \mid \mathcal{F}_{t_1\wedge\tau_n}] \\
&= \mathbb{E}[B_{t_1\wedge\tau_n}^2 \mid \mathcal{F}_{t_1\wedge\tau_n}] \\
&= \mathbb{E}[B_{t_1\wedge\tau_n}^2 - t_1\wedge\tau_n + t_1\wedge\tau_n \mid \mathcal{F}_{t_1\wedge\tau_n}].
\end{aligned}$$

Therefore we have

$$\mathbb{E}[B_{t_1\wedge\tau_n}B_{t_2\wedge\tau_n} - t_1\wedge\tau_n \mid \mathcal{F}_{t_1\wedge\tau_n}] = 0,$$

and

$$\begin{aligned}
\mathbb{E}[|Y_{t_2\wedge\tau_n} - Y_{t_1\wedge\tau_n}|^2] &= \mathbb{E}\left[\sum_{i=1}^{d}\left(\int_{t_1\wedge\tau_n}^{t_2\wedge\tau_n}\sigma_i^n(s)\,dB_s^i\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\sum_{i=1}^{d}\int_{t_1\wedge\tau_n}^{t_2\wedge\tau_n}(B_{s\wedge\tau_n}^i - B_{t_1\wedge\tau_n}^i)\,b_i^n(s)\,ds\right] \\
&\quad + \mathbb{E}\left[\sum_{i=1}^{d}\int_{t_1\wedge\tau_n}^{t_2\wedge\tau_n}|b_i^n(s)|^2\,ds\right] \to 0
\end{aligned}$$

from the proof of (9.3.4). Hence $Y_{t\wedge\tau_n}$ converges to 0 in L^2 and in probability, where $\|b^n\| \to 0$. Since $|Y_t| \ge |Y_t|\mathbf{1}_{\{\tau_n > t\}}$, we can show as in Part 2 that $Y_t = 0$ in probability when $\|b^n\| \to 0$.

9.3.2 Corollary. In particular we have

$$B_t^2 = \int_0^t 2B_s\,dB_s + t.$$

(We already know that $B_t^2 - t$ is a martingale, Itô's Formula gives us the exact form of the martingale.)

exercises

1. Let $\alpha(t,x)$ and $\beta(t,x)$ be two continuous, mean reverting processes.
   (a) Show that the process $X_t = f(x) + \int_0^t \alpha(s,x)ds + \int_0^t \beta(s,x)dW_s$ is continuous and mean reverting.
   (b) Show that if the function $f(x)$ is C^2, but is continuous except at $x=0$, then we have, for each t,

$$E[f(X_t)] = f(X_0) + \int_0^t f'(X_s)\alpha(s,X_s)ds + \frac{1}{2}\int_0^t f''(X_s)\beta^2(s,X_s)ds$$
$$+ \frac{1}{2}\int_0^t f'(0)\beta^2(s,0)ds \quad \text{a.s.}$$

(Hint: use an argument similar to the proof of 9.1.1)

9.9 Bibliography

We have proved Itô's formula in its simplest form. For generalizations and applications to stochastic differential equations one can consult Arnold (1973), McKean (1969), Itô and McKean (1965) and the chapter on diffusions in Gikhman and Skorokhod (1969). A general treatment of martingales and stochastic integrals with respect to martingales is given in Dellacherie and Meyer (1978).

APPENDICES

Appendix 1 — The Symmetric Random Walk in $\mathbb{R}^k$ is Recurrent. Analogous for $k = 1$ and $k = 2$; Transient for $k \geq 3$

Consider a Markov chain with transition probabilities p_{ij}, and let $p_{ij}^{(n)}$ be the probability starting from state i, that the process will be in state j at time n, that is after n transitions. Let $f_i^{(n)}$ be the probability, starting from i, that the first return to i will occur at time n. We then have

$$p_{ii}^{(n)} = \sum_{r=1}^{n} f_i^{(r)} p_{ii}^{(n-r)}.$$

A1.2 *Theorem.*

(a) $\displaystyle\sum_{n=1}^{N} f_i^{(n)} \leq \sum_{n=1}^{N} p_{ii}^{(n)} \sum_{n=1}^{N} p_{ii}^{(n)}$

(b) State i is recurrent if and only if $\sum_{n} p_{ii}^{(n)} = \infty$.

Proof.

(a) $$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N} \sum_{r=1}^{n} f_i^{(r)} p_{ii}^{(n-r)} = \sum_{r=1}^{N} f_i^{(r)} \sum_{n=r}^{N} p_{ii}^{(n-r)}$$

$$\leq \sum_{r=1}^{N} f_i^{(r)} \sum_{n=1}^{N} p_{ii}^{(n)}$$

(b) If $\sum_n p_{ii}^{(n)}$ is not finite, by the Borel–Cantelli Lemma, i is visited only finitely many times and therefore must be transient. If the series diverges, let

f_n is the probability, starting from 0, that there will ever be a return to 0. Then

$$f_n = \sum_{k=1}^{n} f_k^{(n)} = \sum_{k=1}^{n} p_k^{(n)} = \frac{\sum_{k=1}^{n} p_k^{(n)}}{\sum_{k=0}^{n} p_k^{(n)}}$$

by part (a) and the fraction approaches 1 as $N \to \infty$. The claim is now proved. □

The two-dimensional random walk in $\mathbb{R}^2$ is a Markov chain whose state space consists of 2-tuples of integers. At each transition, exactly one coordinate changes, by $+1$ or -1, each with probability $\frac{1}{2}$. The probability of moving from $(s_1, \ldots, s_k)$ to $(s_1, \ldots, s_i + 1, \ldots, s_k)$ as well as the probability of moving from $(s_1, \ldots, s_i, \ldots, s_k)$ to $(s_1, \ldots, s_i - 1, \ldots, s_k)$ is $\frac{1}{2k}$ for $i = 1, \ldots, k$. Thus in one dimension, we have a particle that moves right or left with equal probability. In two dimensions the particle moves right, left, up or down with probability $\frac{1}{4}$ in each case. In three dimensions, a particle at (x, y, z) can move to $(x-1, y, z)$, $(x+1, y, z)$, $(x, y-1, z)$, $(x, y+1, z)$, $(x, y, z-1)$, or $(x, y, z+1)$, and each occurs has probability $\frac{1}{6}$.

4.3 Theorem.

(a) *When $k = 1$ we have $p_{2n}^{(2n)} = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}$ and the states are recurrent.*

(b) *When $k = 2$ we have $p_{2n}^{(2n)} = \left(\frac{1}{4}\right)^{2n} \binom{2n}{n}^2$, and again the states are recurrent.*

(c) *When $k \geq 3$, the states are transient.*

Proof. (a) Starting from 0, the particle will be at 0 after $2n$ steps iff it has taken exactly n steps to the right and n steps to the left. Since the n positions where steps to the right occur can be chosen in $\binom{2n}{n}$ ways, the formula for $p_{2n}^{(2n)}$ follows from $\binom{2n}{n} = \frac{(2n)!}{n!n!}$ and by Stirling's formula,

$$(2n)! \sim (2n)^{2n} e^{-2n} \sqrt{4\pi n} \quad \text{and} \quad n! \sim n^n e^{-n} \sqrt{2\pi n}. \quad \text{Then}$$

$$\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}} \quad \text{and therefore} \quad u_0^{(1)} = \frac{1}{\sqrt{\pi}} \sum \frac{1}{\sqrt{n}} = \infty,$$

the states are recurrent.

(b) Starting from 0, the particle will be at height $2n$ after $2n$ steps if for some r between 0 and $2n$ it has taken exactly r steps right, r steps left, n steps up and $n-r$ steps down. Therefore

$$p_0^{(2n)} = \sum_{r=0}^{2n} \frac{(2n)!}{r!r!(n-r)!(n-r)!} \left(\frac{1}{4}\right)^{2n}$$

Multiply and divide by $(n!)^2$ to obtain

$$p_0^{(2n)} = \left(\frac{1}{4}\right)^{2n} \binom{2n}{n} \sum_{r=0}^{2n} \binom{n}{r}\binom{n}{n-r} \quad \text{recall that: } \binom{n}{r} = \binom{n}{n-r}$$

Now in selecting n objects out of $2n$ we can choose r from the first n and $n-r$ from the second, so the summation is simply $\binom{2n}{n}$. This establishes the desired formula for $p_0^{(2n)}$. By part (a), $p_0^{(2n)} \sim 1/\pi n$, and again the series must diverge.

(c) The argument is entirely analogous. We first note that for all $d \geq 3$, $p_0^{(2n)} = O(n^{-d/2})$, that is, for some constant C we have $p_0^{(2n)} \leq C/n^{d/2}$ for sufficiently large n. In a sequence of $2n$ steps, roughly dn will involve a change in the first coordinate, and in the remaining $2(d-1)n$ the steps will change other coordinates. Thus we can think of choosing an n-dimensional walk and then subjecting it to d successive 1- and 2-step returns to the origin. We can then complete the heuristic argument as follows: to show that the walk is transient for all $d \geq 3$. [A formal proof by induction can be completed by combining $p_0^{(2n)}$ as the probability that a symmetric n-dimensional random walk is back at its original state after $2n$ steps, and then showing that

$$p_0^{(2n)} = \sum_{r=0}^{n} \binom{2n}{2r} \binom{2r}{r} \left(\frac{1}{2}\right)^{2n} p_0^{(2r)} p_0^{(2n-2r)} \leq \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} p_0^{(n)} p_0^{(n)}$$

(There must be $2r$ steps in which the first coordinate changes and it is back at its original value, and $2n-2r$ steps in which the first coordinate changes and it is back at its original value.)]

As we have seen in the proof of Theorem A1.2, $p_0^{(2n)} \sim 1/\sqrt{\pi n}$ for the 1-dimensional walk, and for the d-dimensional walk we have by similar reasoning:

$$p_0^{(2n)} = O(n^{-d/2}) + O(n^{-(d-1)/2}) + \cdots + O(n^{-1})$$

Thus for the d-dimensional walk,

$$p_0^{(2n)} = O(n^{-d/2}) + O(n^{-(d-1)/2}) + \cdots + O(n^{-1})$$

Since $\sum n^{-(d-1)/2}$ converges for $d \geq 3$, the series must converge.

Appendix 2 Semicontinuous Functions

Let $f_1, f_2, \ldots$ be continuous maps from the metric space Ω to the extended reals $\mathbb{R}$, and suppose $f_n(x)$ converges to a limit $f(x)$ for each x. f need not be continuous, however, even if every f_n is. Functions of this type play an important role in many aspects of analysis and probability.

A2.1 *Definition.* Let Ω be a metric space. The function $f: \Omega \to \mathbb{R}$ is said to be *lower semicontinuous* (LSC) on Ω if $f(x) \le \liminf_n f(x_n)$ for every sequence $x_n \to x$ in Ω. Similarly, f is *upper semicontinuous* (USC) on Ω if $f(x) \ge \limsup_n f(x_n)$ for every $x_n \to x$. Thus f is LSC if $-f$ is USC. Note that f is continuous if it is both LSC and USC.

We have the following criterion for semicontinuity.

A2.2 *Theorem.* The function f is LSC on Ω iff, for each sequence $\{x_n\}$ converging to a point $x \in \Omega$, we have $\liminf_n f(x_n) \ge f(x)$, where $\liminf_n f(x_n) = \sup_n \inf_{k \ge n} f(x_k)$. Hence f is USC iff $\limsup_n f(x_n) \le f(x)$ whenever $x_n \to x$.

Proof. Let f be LSC. If $x_n \to x$ and $\liminf_n f(x_n) < f(x)$, then for an infinite subset of $\mathbb{N}$ (hence eventually) $x_n \in f^{-1}(-\infty, r]$, that is, $f(x_n) \le r$ eventually. Thus $\liminf_n f(x_n) \le r < f(x)$. Conversely if $x_n \to x$ implies

$$\liminf_n f(x_n) \ge f(x)$$

we show that $V = \{x : f(x) > a\}$ is open. Let $x_n \to x$, where $f(x) > a$. Then $\liminf_n f(x_n) > a$, hence $f(x_n) > a$ eventually, that is, $x_n \in V$ eventually. Thus V is open. $\square$

We now prove a few properties of semicontinuous functions.

A2.3 *Theorem.* Let f be LSC on the compact metric space Ω. Then f attains its minimum, that is, $f \ge f(x_0)$ on Ω for some $x_0 \in \Omega$. If f is USC on Ω, f attains its supremum.

Proof. If f is not bounded below, there is a sequence $x_n \in \Omega$ with $f(x_n) < -n$. By compactness, we have a subsequence x_{n_k} converging to some $x \in \Omega$. Since f is LSC, $\liminf_k f(x_{n_k}) \ge f(x)$. But $f(x_{n_k}) < -n_k \to -\infty$, so that $f(x) = -\infty$, contrary to $f(x) > -\infty$. $\square$

A2.4 *Theorem.* Let f_α be LSC on Ω for each $\alpha \in A$, then $\sup_\alpha f_\alpha$ is LSC on Ω. Hence $\inf_\alpha f_\alpha$ is USC. (Hence if f_α is USC on each Ω, then $\inf_\alpha f_\alpha$ is USC, and if A is finite, then $\min_\alpha f_\alpha$ is USC.)

Proof. Let $f = \sup\{g : g \text{ is USC and } g \le f_1, \dots, f_n\}$. Then $f \le f_j$ for each j, so $f(x) < \alpha$ for all $x \in U_j$. If $g = \min(f_1, \dots, f_n)$, then

$$\{x : g(x) < \alpha\} = \bigcup_{j=1}^{n} \{x : f_j(x) < \alpha\}$$

is open. ∎

42.5 Theorem. Let $\Gamma : \Omega \to \mathcal{P}(\Omega)$ be a multivalued mapping. Define

$$\underline{f}(x) = \liminf_{y \to x} f(y), \qquad x \in \Omega;$$

that is,

$$\underline{f}(x) = \sup_{\rho > 0} \inf_{y \in B(x,\rho)} f(y),$$

where B ranges over all open balls with center x and radius $1/n$, $n = 1, 2, \dots$. Then $\underline{f}$ is LSC on Ω and $\underline{f} \le f$; furthermore, if g is LSC on Ω and $g \le f$, then $g \le \underline{f}$. Thus $\underline{f}$, called the lower envelope of f, is the sup of all LSC functions that are less than or equal to f; there is always a largest such function, namely the function identically $-\infty$.

Similarly, if $\overline{f}(x) = \limsup_{y \to x} f(y) = \inf_{\rho > 0} \sup_{y \in B(x,\rho)} f(y)$, then $\overline{f}$, the upper envelope of f, is USC and $\overline{f} \ge f$, in fact $\overline{f}$ is the inf of all USC functions that are greater than or equal to f.

Proof. It suffices to consider $\underline{f}$. Let $\{x_n\}$ be a sequence in Ω with $x_n \to x$ and $\liminf_{n \to \infty} \underline{f}(x_n) = \beta < \underline{f}(x)$. If V is a neighborhood of x, we can choose a ball B such that $x \in B \subset V$ and $\overline{f}(x) > \beta$. Since V is also a neighborhood of x_n, we have

$$\beta < \underline{f}(x) \le \inf_{y \in V} f(y),$$

so

$$\underline{f}(x) = \sup_{V} \inf_{y \in V} f(y) \le \beta < \underline{f}(x).$$

a contradiction. By 42.2, $\underline{f}$ is LSC, and $\underline{f} \le f$ by definition of $\underline{f}$. Finally if g is LSC, $g \le f$, then $g(x) = \liminf_{y \to x} g(y) \le \liminf_{y \to x} f(y) = \underline{f}(x)$ since g is LSC. To prove $\inf_{y \in B} f(y) < \underline{f}(x) + \varepsilon$ for given ε and each B, pick $x_B \in B$ with $f(x_B) < \underline{f}(x) + \varepsilon$. If we do this for $B = B_n = B(x, 1/n)$, $n = 1, 2, \dots$, we get $x_n \to x$. Now a sequence consisting of x_n 's will fail to satisfy $\liminf_{n \to \infty} \underline{f}(x_n) < \underline{f}(x)$, so given $\varepsilon > 0$, $f(x_n) < \underline{f}(x) + \varepsilon$ for infinitely many n. ∎

12.6 *Theorem.* Let Ω be a metric space. For LSC functions on Ω there is a sequence of continuous functions $f_n : \Omega \to \mathbb{R}$ such that $f_1 \le f_2 \le \cdots$. Thus if f is LSC, there is a sequence of continuous functions $f_n \le f$ with $f_n \uparrow f$. If $f \ge 0$ the f_n may be chosen so that $f_n \ge 0$ for all n.

PROOF. (Following Tornheim, 1964.) First assume $f \ge 0$ and drop value ∞. For each n define $g(x) = \inf\{f(y) : y \in \Omega\}$. Let $x, y \in \Omega$, where $x \ne y$. Then $g(x) \le f(y) \le g(y) + d(x, y)$. Since $g(x) \le f(x)$ and $g(y) \le f(y)$ we have $g(x) \le g(y) + d(x, y)$.

Let $x, y \in \Omega$. Then $g(x) \le \inf\{d(x, z) + f(z) : z \in \Omega\} \le d(x, y) + \inf\{d(y, z) + f(z) : z \in \Omega\} = d(x, y) + g(y)$. Take the infimum over z in addition $g(x) \le g(y) + d(x, y)$. By symmetry,

$$|g(x) - g(y)| \le d(x, y).$$

Hence g is continuous on Ω.

Now set $f_n(x) = \inf\{f(z) + n d(x, z) : z \in \Omega\}$. Then $0 \le f_n(x) \le f(x)$. But given $\varepsilon > 0$, for each x we can choose $z_x \in \Omega$ such that

$$f_n(x) + \varepsilon > f(z_x) + n d(x, z_x) \ge \inf\{f(z) : z \in \Omega\}.$$

But $f_n(x) + \varepsilon \ge f(z_x)$, and it follows that $d(x, z_x) < \varepsilon$. Since f is LSC, for $\varepsilon > 0$ and $x \in \Omega$, there is $\delta > 0$ such that $f(z) > f(x) - \varepsilon$ whenever $d(x, z) < \delta$. But now

$$f_n(x) = \inf\{f(z) + n d(x, z) : z \in \Omega\} \ge \inf\{f(z) : z \in \Omega\} > f(x) - \varepsilon$$
$$\ge f(x) - 2\varepsilon \qquad \text{for } n \text{ sufficiently large.}$$

Thus we have shown that $0 \le f_n \le f$ and $f_n \uparrow f$. If f is LSC and not necessarily nonnegative, if $f \ge g$ where g is continuous, then $f_n \ge g$ and $f_n \uparrow f$.

In general, let $h(x) = 2\pi^{-1/2} \int_{-\infty}^{x} e^{-t^2}\,dt$. Then h is an order preserving homeomorphism of $\mathbb{R}$ onto $(0, 1)$. If f is LSC, then $h \circ f$ is finite-valued, LSC, and nonnegative, so there are continuous functions g_n such that $g_n \uparrow h \circ f$. Then $h^{-1} \circ g_n \uparrow f$. ∎

Appendix 3 Construction of the Proof of Theorem 7.3.2

The "if" part of the theorem requires no proof. We may assume without loss of generality that $\int X_n = 0$. The said condition is equivalent to the statement that, for each $\varepsilon > 0$, there is $n = n(\varepsilon)$ such that

$$\sup_{\|a\|=1} \left| 1 - h\left(\begin{pmatrix} a \\ \cdot \\ \cdot \end{pmatrix} \right) \right| < \varepsilon \qquad \text{for } n \ge n(\varepsilon).$$

Furthermore, in this case

$$\max_{|z|=R} |\quad| \to 0$$

uniformly for α in any compact in small case Problem 3, Section 7.3).

(a) We show first that

$$\quad \to 0 \qquad \text{as} \qquad \to \infty,$$

Let ψ denote the characteristic function of the random variable X. By the inversion formula,

$$\quad = \quad \to \quad$$

and the convergence is uniform for α in any compact interval $[\alpha_1, \alpha_2]$. If "Log" denotes the principal branch of the logarithm, then

$$\quad$$

is the uniformly unique continuous logarithm of ψ_n by having the value 0 at $\alpha = 0$. Thus

$$\quad$$

Therefore,

$$\quad \to 0$$

where each θ_k is an integer. But $|\psi_n| \to |\psi| = 1$ near $\alpha = 0$, so $\limsup_{n\to\infty} |\theta_n| = \limsup_{n\to\infty} |\theta_n - \theta_{n-1}|$. Essentially $\theta_n - \theta_{n-1} \to 0$, and then $\limsup_{n\to\infty} |\theta_n| \le \limsup_{n\to\infty} |\theta_n - \theta_{n-1}|$; hence

$$\limsup_{n} \frac{n^2}{2} + \sum_{k=1}^{n} \left[\hat{a}_k \left(\frac{k}{n} \right) - 1 \right]$$

$$- \liminf_{n} \sum_{k=1}^{n} \alpha_k \left[\hat{a}_k \left(\frac{k}{n} \right) \right]^2$$

$$\leq \limsup_{n} \max_{1 \leq k \leq n} \left| \hat{a}_k \left(\frac{k}{n} \right) - \sum_{j=1}^{k} \hat{a}_j \left(\frac{k}{n} \right) - 1 \right|$$

Now

$$\max_{1 \leq k \leq n} \hat{a}_k \left(\frac{k}{n} \right) - 1 \Big| \to 0$$

and

$$\sum_{k=1}^{n} \left| \hat{a}_k \left(\frac{k}{n} \right) - 1 \right| = \sum_{k=1}^{n} \left| \int_0^1 \cos \left(\frac{\pi k u}{n} \right) - 1 \, dW(u) \right|$$

$$\leq \sum_{k=1}^{n} \left| \int_0^{\frac{1}{n}} \cos \left(\frac{\pi k u}{n} \right) dW(u) \right|$$

$$\leq \frac{\pi^2}{2n} \sum_{k=1}^{n} k \to \frac{\pi^2}{2} < \infty.$$

This proves (a).

(b) To show that

$$\limsup_{n \to \infty} \left| \frac{1}{n} - \sum_{k=1}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \hat{a}_k(u) \, dW(u) \right| < \frac{4}{\pi^2 n^2}$$

first by (a),

$$\frac{n^2}{2} - \sum_{k=1}^{n} \left[\hat{a}_k \left(\frac{k}{n} \right) - 1 \right] \to 0,$$

that is,

$$\frac{n^2}{2} - \sum_{k=1}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left(1 - \cos \frac{\pi k}{n} \right) dW(u) \to 0$$

or equivalently,

$$\frac{n^2}{2} - \sum_{k=1}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left(1 - \cos \frac{\pi k}{n} \right) dW(u)$$

$$- \sum_{k=n+1}^{\infty} \int_{\frac{k-1}{n}}^{\frac{k}{n}} \left(1 - \cos \frac{\pi k}{n} \right) dW(u) \to 0$$

Again, noting that $\sin\theta = 0$ for $\theta = k\pi$, $k = \pm 1, \pm 2, \ldots$, and $\sin\theta \neq 0$ for $\theta \neq k\pi$, we have

$$\limsup_{n\to\infty} \left| \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dF(x) \right|$$

$$= \limsup_{n\to\infty} \left| \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dH_n(x) \right| . \qquad (1)$$

But

$$\sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dF(x) \le \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \frac{n^2x^2}{2\sigma_n^2} dH_n(x)$$

$$\le \frac{n^2}{2\sigma_n^2} \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} x^2 \, dH_n(x) \le \frac{n^2}{2\sigma_n^2} \varepsilon^2$$

and take the absolute values around the left-hand side. Letting $\varepsilon \to 0$ gives the desired result. This also shows that the quantity inside the braces in (1) is nonnegative. Consequently,

$$\limsup_{n\to\infty} \left| \sum_{j=1}^{k} \left| \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dF(x) \right| \right|$$

$$\le \limsup_{n\to\infty} \left| \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dF(x) \right|$$

$$= \limsup_{n\to\infty} \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dH_n(x) \qquad \text{by (1)}.$$

But

$$\sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} \left(1 - \cos\frac{nx}{\sigma_n}\right) dH_n(x) \le 2 \sum_{j=1}^{k} \int_{(a_j, a_j+\varepsilon]} dH_n(x)$$

$$= 2 \sum_{j=1}^{k} P\{a_j < X_n \le a_j + \varepsilon\}$$

$$\le 2 \sum_{j=1}^{k} \frac{\sigma_n^2}{n^2 \varepsilon^2} \qquad \text{by Chebyshev's inequality}$$

$$= \frac{2k}{n^2}, \qquad \text{proving (ii)}.$$

The proof of the theorem is completed by letting $n \to \infty$ in (ii). ∎

Appendix A Proofs of the Convergence of Types Theorem 5.2.4

(a) Let G_n, G, F_n, F be the distribution functions of Z_n, Z, Y_n, Y, respectively.

We may select convergent subsequences $\{a_{n_j}\}, \{b_{n_j}\}$ such that

$$a_{n_j} \to a, \qquad b_{n_j} \to b, \qquad \text{where} \qquad 0 \le a < \infty \qquad -\infty < b < \infty.$$

We first show that $a > 0$.

Suppose that $a = 0$. Let $E = \{x \in \mathbb{R} : \liminf_{j\to\infty} a_{n_j} x + b_{n_j} > -\infty\}$ and $x_0 = \sup E$. (Take $x_0 = -\infty$ if $E = \emptyset$.)

(i) If $x < x_0$, then $a_{n_j} x + b_{n_j} \to -\infty$.

Proof. Let $x < x_0$. Then $x < x_1$ for some $x_1 \in E$. Thus $\liminf_{j\to\infty} a_{n_j} x_1 + b_{n_j} > -\infty$ and $x < x_1$ implies $x \in E$ by definition of E.

Now

$$a_{n_j} x + b_{n_j} = a_{n_j}(x - x_1) + a_{n_j} x_1 + b_{n_j} \to -\infty + \liminf_{k\to\infty} a_{n_k} x_1 + b_{n_k} = -\infty$$

since $x - x_1 < 0$ and $a_{n_j}(x - x_1) \to -\infty$ since $a_{n_j} > 0$ and $x - x_1 < 0$. This proves (i).

(ii) If $x > x_0$, then $G(x) = 0$.

Proof. Let $x > x_0$. Choose $x_1 \in E$ such that $x_1 < x$ and x is a continuity point of G. Then $F_{n_j}(x) \to F(x)$ or eventually $F_{n_j}(x) > 0$. By (i) eventually $a_{n_j} x + b_{n_j} < x_1$ so

$$G_{n_j}(x) = F_{n_j}(a_{n_j} x + b_{n_j}) \to F(x_1) = 0$$

for some subsequence. Thus $G(x) = 0$ or $x \ge x_0$. Let $x > x_0$ be a continuity point of G. Then $G(x) = \lim_{n\to\infty} G_n(x) = 0$ and since $G(x) \ge G(x_0) = 0$, we have $G(x) = 0$, proving (ii).

(iii) If $x < x_0$, then $G(x) = 1$.

Proof. By definition of x_0, we can find a subsequence $\{j\}$ such that $a_{n_j} x + b_{n_j} \to -\infty$. Choose x_1 a continuity point of F such that $F(x_1) > 1 - \varepsilon$. Eventually, $a_{n_j} x + b_{n_j} < x_1$ so $G_{n_j}(x) = F_{n_j}(a_{n_j} x + b_{n_j}) > F(x_1) > 1 - \varepsilon$. Thus $G(x) = 1$ for $x < x_0$. If $x < x_0$ is a continuity point of G we have $G(x) \ge G(x_0) = \lim_{j\to\infty} G_{n_j}(x) = 1$. Proving (iii)

It follows from (ii) and (iii) that G is degenerate, a contradiction.

Next we show that λ is finite.

If $\lambda_n \to \infty$ then $\alpha_n x + \beta_n \to \infty$ for every $x > 0$. The same argument as above may be repeated to show $G(x) = 1$, a contradiction. If $\alpha_n \to 0$, then $\alpha_n x + \beta_n \to \beta$ for each $x \in \mathbb{R}$, so the argument of (i) shows $G(x) = 1$, a contradiction.

Next we show that $\alpha > 0$.

Let c be a continuity point of G, and let $x_1 > 0$ be such that $\alpha x + \beta + x_1$ and $\alpha x + \beta - x_1$ are continuity points of F. Then $\alpha_n x + \beta_n - \alpha_n - b_n \to \alpha x + \beta - \alpha - \beta$, so that $\alpha_n x + \beta_n - \alpha_n - b_n + \alpha_n x + \beta_n - \alpha_n - b_n \to 2\alpha x + 2\beta - 2\alpha - 2\beta$, hence

$$F_n(\alpha_n x + \beta - \alpha_n - b_n) - F_n(\alpha_n x + \beta_n - \alpha_n - b_n) \to G(x) - G(x) = 0.$$

Now we can choose n so that $F(\alpha x + \beta - \alpha - \beta) \le G(x) \le F(\alpha x + \beta - \alpha - \beta)$, since c_1 and c_2 may be chosen arbitrarily close to $G(x)$. Since G is continuously positive on $\mathbb{R}$.

Thus $\alpha > 0$, then $F(x) = G(\alpha x + \beta)$ for all continuity points, and hence for all x. Thus $\alpha = 1$, $\beta = 0$ because $F(x) = 1$, $F(-x) = 0$. Thus α is degenerate at β, a contradiction.

Finally, if G is continuous at x and F is continuous at $\alpha x + \beta$, we have just shown $F(\alpha x + \beta) = G(x) = F(\alpha x + \beta)$, or $G(x) = F(\alpha x + \beta)$. Since there are only countably many real numbers y such that y is discontinuous at y or $\frac{y - \beta}{\alpha}$ is discontinuous at y, it follows that $G(x) = F(\alpha x + \beta)$ for all $x \in \mathbb{R}$.

Now if we have two non-degenerate subsequences $\alpha_n \to \alpha'$, $\beta_n \to \beta'$, the above argument shows that $G(x) = F(\alpha' x + \beta')$ and $G(x) = F(\alpha x + \beta)$, so that $F(\alpha' x + \beta') = F(\alpha x + \beta)$ for all $x \in \mathbb{R}$, so that $\alpha' = \alpha$, $\beta' = \beta$. Random variables with the same distribution have the same characteristic function, therefore

$$\exp\left[\frac{-t^2\sigma^2}{n}\right] \varphi_n\left(\frac{t}{\sigma n}\right) = \exp\left[\frac{-t^2}{n}\right] \varphi_Y\left(\frac{t}{\sigma n}\right) \qquad \text{for all} \quad t.$$

Say $\sigma = \sigma^2$, and set $\varphi = \varphi(\sigma)$. Let $\psi = \psi(\sigma)$ so that

$$\lambda(t) = \left| \varphi_Y\left(\frac{t}{\sigma n}\right) \right| = |\varphi(t/\sigma)|$$

for $\sigma > 0$. Then

$$|\varphi(t)| = \lambda(\sigma t) = |\varphi(\sigma^2 t)| = \cdots = \lambda(\sigma^{2^n} t)| \to |\varphi(0)| = 1$$

It follows that φ is degenerate, a contradiction (see Problem 4.1). Thus $\sigma = \sigma^2 = \sigma^{2^2} = \sigma^{2^3} = \cdots$ is of sufficiently small or large σ. There exist $n_k \to \infty$, $b_{n_k} \to b$, implying $\varphi(t) = \varphi(t)$

Comment. If $a = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n])$ and $A = (\mathrm{cov}[X_i, X_j])$, where a and A are known but not necessarily positive, then if $\alpha = (\alpha_1, \ldots, \alpha_n)$ and we set $\lambda = \alpha A \alpha'$, we obtain a contradiction as above. We can conclude that $\alpha A \alpha' \geq 0$ for all $\alpha \in \mathbb{R}^n$.

(b) (i) Let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then X_α and Y_α are of the same positive type, and if $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, $Y_\alpha \stackrel{L}{=} -\alpha_1 X_1 - \cdots - \alpha_n X_n + \beta + i\gamma$, hence X_α and $-Y_\alpha$ must be of the same positive type.

Let $S = \{\alpha : \alpha_1, \ldots, \alpha_n \in \mathbb{C}, \Delta_\alpha = |\alpha_1 \alpha_2| > 0\}$. If $S = \emptyset$ a suitable part (a) shows that X and $a + bY$ must have the same distribution for some real a, b, $b > 0$ and

$$\lim_{t \to \infty} \varphi_X(t) = 0 \qquad \lim_{t \to \infty} \varphi_Y(t) = 0$$

Now suppose that $S \neq \emptyset$. Then $\varphi_X = \varphi_Y$ implies $\varphi_X^2 = \varphi_Y^2$ (see 5.2.9), and it follows from (a) that for some real a, V with $V > 0$, we have $X \stackrel{L}{=} a + V + Y$, so

$$\lim_{t \to \infty} \varphi_X(t) = e^{ia} \qquad \lim_{t \to \infty} \varphi_Y(t) = 0$$

Now there are three possibilities.

Case 1. X and Y are both infinite. Then since $\varphi^2(t) = \lim_{n \to \infty} \varphi(t)^{2n} \{P = 1\} + \lim_{n \to \infty} X_{2n}$, we have $|\varphi(t)| = |\varphi(t)|^2$ (use the convention that $0^0 = 0$ here). Thus $|\varphi(t)| = 0$ or 1 and the result follows.

Case 2. X is infinite, Y finite. Then $\varphi_X(t) = e^{ia}$, $\varphi_Y(t) = 0$, and $X \stackrel{L}{=} a + Y$ $\{P = 1\}$, proving the result.

Case 3. X is finite. As before, there are a, b, $b > 0$, $\varphi_X = e^{ia}$, and $X \stackrel{L}{=} a + bY$ $\{P = 1\}$, and the result follows.

Appendix 5 The Multivariate Normal Distribution

In this appendix we will denote a column vector with components $x_1, \ldots, x_n$ by x, a row vector with components $\lambda_1, \ldots, \lambda_n$ by λ, and X a random (column) vector with components $X_1, \ldots, X_n$. The superscript t will indicate the transpose of a matrix. To avoid awkward special cases, we agree that normal with mean μ and variance $\sigma^2 = 0$ means degenerate at μ.

A5.1 *Definition.* The random variables $X_1, \ldots, X_n$ are said to be *jointly Gaussian* or the random vector $X = (X_1, \ldots, X_n)$ is said to be *Gaussian* if the characteristic function of X is

$$\varphi(\lambda_1, \ldots, \lambda_n) = \exp\{i\lambda b' - \tfrac{1}{2} \lambda A \lambda'\} \qquad (*)$$

where the b_j are arbitrary real numbers and A is a symmetric, nonnegative definite matrix with real entries.

A much more concise interpretation is possible.

A.5.2 *Theorem.* The random vector X is Gaussian if and only if X can be expressed as $AY + b$ where the Y_j are independent normal random variables with 0 mean.

Proof. If $X = AY + b$, then $E \exp[i\theta' X] = \exp[i\theta' b] E \exp[i\theta' AY]$. But

$$E \exp[i\theta' Y] = E \prod_{j=1}^{n} \exp[i\theta_j Y_j] = \prod_{j=1}^{n} E \exp[i\theta_j Y_j]$$

$$= \exp\left[\frac{1}{2} \sum_{j=1}^{n} \lambda_j \theta_j^2 \right] = \exp\left[-\frac{1}{2} \theta' D \theta \right]$$

where D is a diagonal matrix whose entries are $\lambda_1, \ldots, \lambda_n$. Now $E[Y_j] = 0$, $j = 1, \ldots, n$. So $Y = AY + b$ has a characteristic function of X given by

$$E[\exp(i\theta' X)] = \exp[i\theta' b - \tfrac{1}{2} \theta' A D A' \theta]$$

where $A = A(Y)$. Since the diagonal matrix D is symmetric, so is A, and A is nonnegative definite as well since $\theta' A \theta = \theta' B B' \theta = \sum_{j=1}^{n} (\theta' B_j)^2 \geq 0$.

Conversely, assume that X is Gaussian with characteristic function given by (1) above. Let A be an orthogonal matrix such that $A' \Sigma A = D$, a diagonal matrix whose entries are the eigenvalues λ_j of Σ. Let

$$b = E(X), \quad Y = A'(X - b).$$

Then

$$E[\exp(i\theta' Y)] = \exp[-i\theta' A' b] E[\exp(i\theta' A' X)]$$

$$= \exp[-i\theta' A' b] \exp[i\theta' A' b - \tfrac{1}{2} \theta' A' \Sigma A \theta]$$

$$= \exp[-\tfrac{1}{2} \theta' D \theta] \quad \text{where} \quad D = A' \Sigma A.$$

Thus

$$E[\exp(i\theta' Y)] = \exp\left[-\frac{1}{2} \theta' A' \Sigma A \theta \right] = \exp\left[-\frac{1}{2} \theta' D \theta \right]$$

$$= \exp\left[\frac{1}{2} \sum_{j=1}^{n} \lambda_j \theta_j^2 \right].$$

The form of the characteristic function $\varphi(t)$ shows that $Y_1, \ldots, Y_n$ are independent, and Y_i is normal with mean 0 and variance λ_i. Since A is orthogonal, we have $A^{-1} = A^T$, so that $X = AY + b$. ∎

A.5.2 Corollary. *If Σ is any symmetric nonnegative definite matrix, then there is always a Gaussian vector X whose characteristic function is given by A.5.1 with the prescribed b and K.*

Proof. Let A be an orthogonal matrix with $A^T K A = D$, the diagonal matrix of eigenvalues of K, and $X = AY + b$, where the Y_i are independent and normal $(0, \lambda_i)$. The first part of the proof of Theorem A.5.2 shows that the characteristic function of X is of the desired form. ∎

We may give a probabilistic interpretation of the vector b and the matrix K.

A.5.3 Theorem. *If X is Gaussian with characteristic function given by A.5.1, then $E(X) = b$; in other words, b_i is the expectation of X_i, $i = 1, \ldots, n$. Furthermore, K is the covariance matrix of the X_i, that is, $K_{ij} = \mathrm{Cov}(X_i, X_j)$ for all i, j.*

Proof. Let $X = AY + b$ as in Theorem A.5.2. Since the Y_i have finite second moments, so do the X_i. By linearity of the expectation we have $E(X) = b$. $\Sigma(X)$ is the matrix whose ij entry is $E[(X_i - b_i)(X_j - b_j)]$. Thus the covariance matrix of the X_i can be written as

$$\Sigma(X) = E[(X - b)(X - b)^T] = E[(AY)(AY)^T] = AE[YY^T]A^T = ADA^T,$$

where D is a diagonal matrix with entries $\lambda_i = \mathrm{Var}\, Y_i$. But by the last part of the proof of Theorem A.5.2, $ADA^T = K$. ∎

We now show that if the covariance matrix is K is nonsingular, then X has a density.

A.5.4 Theorem. *Let X be Gaussian with mean vector b and covariance matrix K. If K is nonsingular, then the $X_i - b_i$ are mutually independent if and only if $\sum_{i=1}^{n} c_i(X_i - b_i) = 0$ a.s., $(c_1, \ldots, c_n) \neq 0$, then $c_1 = \cdots = c_n = 0$ for all c_i.*

Furthermore, X has density f_X, where

$$f_X(x) = (2\pi)^{-n/2} (\det K)^{-1/2} \exp\left[-\tfrac{1}{2}(x - b)^T K^{-1}(x - b)\right]$$

Proof. Let A be an orthogonal matrix such that $A'\Sigma A = D$, and let $X = AY + b$ as in Theorem 4.5.1. If X is nonsingular, then every eigenvalue λ_i is strictly positive, so Y has density g, where

$$g(y) = |\det A|\, f(Ay + b)\, |\det D|^{1/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \frac{y_i^2}{\lambda_i}\right]$$

$$= (2\pi)^{n/2} f(Ay + b) \exp\left[-\frac{1}{2}(y - \beta)'\Sigma^{-1}(y - \beta)\right].$$

Now the Jacobian of the transformation $x = Ay + b$ is not A, which is ± 1 because A is orthogonal. Since $g = f(x)$, X has density f, where

$$f(x) = (2\pi)^{-n/2} (\det \Sigma)^{-1/2} \exp\left[-\frac{1}{2}(x - b)'A\Sigma^{-1}A'(x - b)\right].$$

But $A\Sigma A' = D$, and it follows that $\Sigma^{-1} = AD^{-1}A'$ and therefore $A(D^{-1})A' = \Sigma^{-1}$.

Now if $c'(X - b) = \sum_{i=1}^{n} c_i(X_i - b_i) = 0$ a.s. then

$$0 = E[c'(X - b)]^2 = E[c'(X - b)(X - b)'c]$$

$$c'E(X - b)(X - b)'c = c'\Sigma c.$$

But Σ is nonsingular, and therefore positive definite, so c must be zero for all c.

If X is singular, the last part of the proof of Theorem A5.5 shows that $c'(X - b) = 0$ a.s. for linearly dependent c. For $c'(X - b) = c'A(Y - \beta) = \sum_{i=1}^{n} c_i'(Y_i - \beta_i)$ where $c_i' = c'A_i$. Since at least one λ_i must be 0, we can choose a nonzero c_i and hence a nonzero c, such that $c'(X - b) = 0$ a.s. If, say, $\lambda_1 = 0$, then $Y_1 = \beta_1$ a.s. Let $\{Y_{i_1}, \ldots, Y_{i_k}\}$ be a maximal linearly independent subset of $\{Y_2 = \beta_2, \ldots, Y_n = \beta_n\}$, then $(Y_1, \ldots, Y_n)$ has a density of the form given in Theorem A5.5, where k is determined by the symmetric dimensional by the first k rows and columns of Σ. The remaining random variables $Y_j = \beta_j$, $i = k + 1, \ldots, n$ can be expressed as a set of probability 1 as linear combinations of the $Y_{i_1}, \ldots, Y_{i_k}$.

The result that nonsingularity of X is equivalent to linear independence of the $X_i - b_i$ holds for arbitrary random variables with finite second moments, as the above argument shows.

A5.6 *Theorem.* (a) If X is Gaussian and $Z = CX$ where C is given by a matrix, then Z is Gaussian.

Bibliography

Apostol, T.M. (1957). *Mathematical Analysis*. Reading, MA: Addison-Wesley.

Ash, R. [illegible]. [illegible]. [illegible]: IEEE Press.

Ash, R.B. [illegible]. [illegible]

Ash, R.B. [illegible]. [illegible]. New York: Academic Press.

Ash, R.B. [illegible]. [illegible]. New York: [illegible].

Ash, R.B. [illegible]. [illegible]. New York: [illegible]. New York: Dover.

Ash, R.B. and [illegible]. [illegible]. New York: Academic Press.

[illegible]. New York: Academic Press.

Billingsley, P. [illegible]. *Convergence of Probability Measures*. New York: Wiley.

Billingsley, P. [illegible]. *Ergodic Theory and Information*. New York: Wiley.

Breiman, L. [illegible]. *Probability*. Reading, MA: Addison-Wesley.

Breiman, L. [illegible]. The individual ergodic theorem of information theory. *Ann. Math. Statist.* [illegible]

Chung, K.L. [illegible]. [illegible]

Chung, K.L. [illegible]. *A Course in Probability Theory*. New York: [illegible].

Chung, K.L. [illegible]. [illegible]

Cornfeld, I.P., Fomin, S.V., and Sinai, Ya.G. [illegible]. *Ergodic Theory*. New York: Springer-Verlag.

Dellacherie, C. and Meyer, P.A. [illegible]. *Probabilités et Potentiel*. [illegible]: Hermann.

Doob, J.L. [illegible]. *Measure Theory*. New York: Springer-Verlag.

Doob, J.L. [illegible]. *Stochastic Processes*. New York: Wiley.

Dowker, Y.N. [illegible]. [illegible]

Dunford, N. and Schwartz, J.T. [illegible]. *Linear Operators*. New York: Wiley. [illegible]

Feller, W. [illegible]. *An Introduction to Probability Theory and its Applications*, Vol. II. New York: Wiley.

Feller, W. [illegible]. *An Introduction to Probability Theory and its Applications*, Vol. I. New York: Wiley.

Ferguson, T.S. [illegible]. *A Course in Large Sample Theory*. New York: Chapman and Hall.

Folland, G.B. [illegible]. *Real Analysis*. New York: Wiley-Interscience.

Gallager, R.G. [illegible]. *Information Theory and Reliable Communication*. New York: Wiley.

Garsia, A.M. [illegible]. *Topics in Almost Everywhere Convergence*. Chicago: Markham.

Gikhman, I.I. and Skorokhod, A.V. [illegible]. *Introduction to the Theory of Random Processes*. Philadelphia: Saunders.

Gnedenko, B.V. and Kolmogorov, A.N. [illegible]. *Limit Distributions for Sums of Independent Random Variables*. Reading, MA: Addison-Wesley.

Gray, R.M. [illegible]. [illegible]

Halmos, P.R. [illegible]. *Lectures on Ergodic Theory*. Tokyo: Mathematical Society of Japan.

Halmos, P.R. [illegible]. [illegible]. New York: Chelsea.

Halmos, P.R. [illegible]. *Measure Theory*. Princeton, NJ: Van Nostrand Reinhold.

[illegible]. [illegible]. New York: Chelsea.

Itô, K. and McKean, H.P., Jr. (1965). Diffusion Processes and Their Sample Paths. New York: Academic Press.

Jacobs, K. (1962). Lecture Notes on Ergodic Theory. Aarhus: [illegible]

Karlin, S. (1966). A First Course in Stochastic Processes. New York: Academic Press.

Kemeny, J.G. and Snell, J.L. (1960). Finite Markov Chains. Princeton, NJ: Van Nostrand.

Kemeny, J.G., Snell, J.L., and Knapp, A.W. (1966). Denumerable Markov Chains. Princeton, NJ: Van Nostrand-Reinhold.

Kieffer, J. (1970). A Generalization of the Shannon-McMillan Theorem and Its Applications to Information Theory. Illinois: University of Illinois.

Kolmogorov, A.N. (1959). Entropy per unit time as a metric invariant of automorphisms. Dokl. Akad. Nauk SSSR 124, 754–755.

Kolmogorov, A.N. (1958). A new metric invariant of transitive dynamical systems and automorphisms in Lebesgue spaces. Dokl. Akad. Nauk SSSR 119, 861–864.

Lévy, P. (1937). Théorie de l'addition des variables aléatoires. Paris: Gauthier-Villars.

Lusternik, L. and Sobolev, V. (1961). Elements of Functional Analysis. New York: Ungar.

[illegible] New York: New York University.

McKean, H.P., Jr. (1969). Stochastic Integrals. New York: Academic Press.

McMillan, B. (1953). The basic theorems of information theory. Ann. Math. Statist. 24, 196–219.

Meyer, P.A. (1966). Probability and Potentials. Waltham, Mass: Blaisdell.

Neveu, J. (1965). Mathematical Foundations of the Calculus of Probability. San Francisco: Holden-Day.

Ornstein, D.S. (1974). Ergodic Theory, Randomness, and Dynamical Systems. New Haven: Yale University Press.

Ornstein, D.S. (1970). Bernoulli shifts with the same entropy are isomorphic. Adv. in Math. 4, 337–352.

Ornstein, D.S. (1973). An example of a Kolmogorov automorphism that is not a Bernoulli shift. Adv. in Math. 10, 49–62.

Parthasarathy, K.R. (1967). Probability Measures on Metric Spaces. New York: Academic Press.

[illegible] New York: Wiley.

[illegible]

[illegible] New York: Academic Press.

Royden, H.L. (1968). Real Analysis. New York: Macmillan. 2nd edition (1968).

Rudin, W. (1966). Real and Complex Analysis. New York: McGraw-Hill.

[illegible] New York: Macmillan.

[illegible] New York: Wiley.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623–656. Reprinted in Shannon, C.E. and W. Weaver (1949). The Mathematical Theory of Communication. Urbana, Ill.: University of Illinois Press.

Shields, P.C. (1973). The Theory of Bernoulli Shifts. Chicago: The University Press of Chicago.

Shiryayev, A.N. (1984). Probability. New York: Springer-Verlag.

Taylor, A.E. (1958). Introduction to Functional Analysis. New York: Wiley.

[illegible]

[illegible] New York: Academic Press.

Williams, D. (1991). Probability with Martingales. Cambridge: Cambridge Univ. Press.

[illegible] Cambridge: Cambridge Univ. Press.

Wong, E. (1971). Stochastic Processes in Information and Dynamical Systems. New York: McGraw-Hill.

Yosida, K. (1965). Functional Analysis. New York: Springer-Verlag.

Solutions to Problems

Chapter 1

Section 1.1

1. We have $\limsup_n A_n = (-1, 1]$, $\liminf_n A_n = \{0\}$.
3. Using $\limsup_n A_n = \{x : x \in A_n$ for infinitely many $n\}$, $\liminf_n A_n = \{x : x \in A_n$ for all but finitely many $n\}$, we obtain

$$\limsup_n A_n = \{(x, y) : x^2 + y^2 \leq 1\},$$

$$\liminf_n A_n = \{(x, y) : x^2 + y^2 < 1\} = \{(0, 0)\}.$$

6. If $x \in \limsup_n (a_n, b_n)$, then $\limsup_n A_n \neq \emptyset$, so $x \in [-\infty, \infty)$. If $x \in A_n$ for infinitely many n, then $x < b_n$ for infinitely many n, because $b_n \to a$. Thus $\limsup_n A_n \subset (-\infty, a]$. But $x \in (a_n, b_n)$ for infinitely many n, so $x \geq a$. Thus $\limsup_n A_n = [a, a]$. Similarly, $\liminf_n A_n = [a, a]$. Thus $(a_n, b_n) \to [a, a]$, and the result follows. The same result is valid for half-open intervals $(a_n, b_n]$ or $[a_n, b_n)$.

Section 1.2

1. (a) If $A = \{-\infty, \infty\}$, then $\mu(\{x\}) = \mu(\{x\} \cap \{\infty\}) + \mu(\{x\} \cap \{-\infty\}) = 0$, and $\mu(\{\infty\}) = \mu(\{-\infty\}) = \infty$. Finite additivity follows quickly. If $A = \{-\infty, a\}$, then $A_n \uparrow -\infty$, but $\mu(A_n) = \infty \not\to \mu(A) = 0$. Thus μ is not continuous from below. Hence it is not countably additive.
   (b) Finiteness of μ follows from the definition, since $\mu(-\infty, a] = a - n$ is unbounded.
3. We have $\mu(\bigcup_{i=1}^n A_i) = \mu(\bigcup_{i=1}^n A_i \cap \Omega) = \sum_{i=1}^n \mu(A_i)$ for all n, hence μ is countably additive. Now apply the dominated convergence theorem.
5. The minimal σ-field of $\mathcal{B}(\mathbb{R})$ which is also the minimal field is unrealistic as the collection $\mathcal{B}_0$ of all finite unions of sets of the form $B_1 \cap B_2 \cap \cdots \cap B_n$, where B_i is either A_i or A_i^c. Any σ-field containing $A_1, \ldots, A_n$ must contain

all sets in $\mathscr{S}$, hence $\mathscr{M} \supset \mathscr{S}$. But $\mathscr{S} \supset \mathscr{R}$ since $\mathscr{S}$ is closed under finite intersections. Since $\mathscr{M}$ and $\mathscr{R}$ are σ-rings, $\mathscr{M} \supset \mathscr{R}$. Thus $\mathscr{M} = \mathscr{R}$. Since $\mathscr{R}$ is closed under countable unions, $\mathscr{R} \supset \mathscr{S}$. Since $\mathscr{S}$ is closed under finite intersections, $\mathscr{S} \supset \mathscr{R}$. Thus $\mathscr{S} = \mathscr{R}$.

Now let $\mathscr{C}$ be the collection of all sets $S_1 \cap \cdots \cap S_n$ where n is a positive integer and each S_i is either A_i or $X - A_i$. Since $A_1 \cap \cdots \cap A_n$ is the intersection of X and n sets of the form A_i or $X - A_i$, $\mathscr{C} \subset \mathscr{S}$. Since $\mathscr{S}$ is closed under finite intersections, $\mathscr{S} \supset \mathscr{C}$. Thus $\mathscr{S} = \mathscr{C}$. Since $\mathscr{C}$ is the collection of all sets that can be generated from the A_i's by taking unions, intersections and complements, $\mathscr{C}$ is closed under complementation. Thus $\mathscr{C} = \mathscr{S} = \mathscr{R}$.

4. (a) As in Problem 3, any set in $\mathscr{C}$ is a finite union of sets in $\mathscr{S}$, hence $\mathscr{C} \supset \mathscr{S}$. But $\mathscr{C}$ is closed under finite unions, so $\mathscr{C} \supset \mathscr{R}$. If $A_1 - \bigcup_{i=2}^{n} B_i \in \mathscr{R}$, then $(A_1 - \bigcup_{i=2}^{n} A_i) \cap \bigcap_{i=2}^{n} B_i^c \in \mathscr{R}$, which belongs to $\mathscr{C}$ because of the distributive law $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

(b) Note that the complement of a finite intersection $\bigcap_{i=1}^{n} A_i$ belongs to $\mathscr{C}$; for example if $B_1, B_2 \in \mathscr{C}$, then

$$(B_1 \cap B_2)^c = B_1^c \cup B_2^c$$
$$= (B_1^c \cap B_2) \cup (B_1 \cap B_2^c) \cup (B_1^c \cap B_2^c) \in \mathscr{C}.$$

Now $\mathscr{C}$ is closed under finite intersections by the distributive law, and it follows from this and the above remark that $\mathscr{C}$ is closed under complementation and is therefore a field. Just as in the proof that $\mathscr{S} = \mathscr{R}$, we find that $\mathscr{C} = \mathscr{R}$.

(c) This is immediate from (a) and (b).

5. (a) Let $A_\alpha \in \mathscr{A}$ for $\alpha = 1, 2, \ldots$. Then A_α belongs to some $\mathscr{B}_{\beta(\alpha)}$ and we may assume $\beta(\alpha) \leq \alpha$. Then $A_\alpha \in \mathscr{B}_\beta$ for $\beta = \sup_\alpha \beta(\alpha)$. Thus $\bigcup_\alpha A_\alpha \in \mathscr{B}_\beta$, hence $\bigcup_\alpha A_\alpha \in \mathscr{A}$. Thus $\bigcup_{\alpha<\omega_1} A_\alpha \in \mathscr{A}$ for $\omega = \omega_1$. Now if $\omega < \omega_1$ and $\bigcup_{\alpha<\beta} A_\alpha \in \mathscr{A}$ for $\beta < \omega$, then $\bigcup_{\alpha<\omega+1} A_\alpha$ belongs to some $\mathscr{B}_\gamma$, hence $\bigcup_{\alpha<\omega+1} A_\alpha \in \mathscr{B}_{\gamma+1} \subset \mathscr{A}$.

(b) We have $\operatorname{card} \mathscr{B}_\alpha \leq \aleph_0$ for all α. This is true for $\alpha = 0$ by hypothesis. If it is true for $\alpha < \omega$, then $\bigcup_{\beta<\alpha} \mathscr{B}_\beta$ has cardinality at most $\operatorname{card} \alpha \cdot \aleph_0 = \aleph_0$. Now if $\mathscr{A}$ has cardinality $\aleph_1$, then $\mathscr{A}$ has cardinality at most $\aleph_1^{\aleph_0} = 2^{\aleph_0 \cdot \aleph_1} = 2^{\aleph_1} = \aleph_1$, but $\operatorname{card} \mathscr{A} \geq \aleph_1$. It follows that $\bigcup_{\alpha<\omega_1} \mathscr{B}_\alpha$ has cardinality $\aleph_1$ or greater.

Section A.1

2. (a) Since $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$ we have $\Omega = \sum_{n=0}^{\infty} \Omega_n$ and $\mathscr{F}$ is clearly closed under complementation. If $E_1, E_2, \ldots \in \mathscr{F}$ and $A \subset \Omega$, then

$$\lambda_{\beta,\gamma}(A \cap \bigcup_{i=1}^{\infty} E_i) = \lambda_{\beta,\gamma}[\bigcup_{i=1}^{\infty} (A \cap E_i)]$$
$$= \sum_{i=1}^{\infty} \lambda_{\beta,\gamma}(A \cap E_i) \quad \text{since} \quad A \cap E_i \in \mathscr{F}$$
$$= \lambda_{\beta,\gamma}(A) + \lambda_{\beta,\gamma}(A \cap E_1^c).$$

Thus

$$\begin{aligned}\lambda(A \cap E) + \lambda(A \setminus E) &= \lambda(A \cap E \cap F_1) + \lambda(A \cap E \setminus F_1) \\ &\quad + \lambda(A \setminus E) \qquad \text{since } F_1 \subset E \\ &= \lambda(A \cap F_1) + \lambda(A \setminus F_1) \qquad \text{since } F_1 \in \mathcal{M}.\end{aligned}$$

This proves that F_1 is a set. Also, if E and F are disjoint we have

$$\begin{aligned}\lambda(A \cap (E \cup F)) &= \lambda(A \cap (E \cup F) \cap E) + \lambda(A \cap (E \cup F) \setminus E) \\ &= \lambda(A \cap E) + \lambda(A \setminus E) \\ &= \lambda(A \cap E) + \lambda(A \setminus E) \qquad \text{since } E \in \mathcal{M}.\end{aligned}$$

Now if the E_n are disjoint sets in $\mathcal{M}$ and $A_n = \bigcup_{i=1}^{n} E_i$, then

$$\begin{aligned}\lambda(A_n) &= \lambda(A_n \cap E_n) + \lambda(A_n \setminus E_n) \\ &\quad \text{since } A_n \text{ belongs to } \mathcal{M} \text{ by hypothesis} \\ &= \lambda(A_n \cap E_n) + \lambda(A_{n-1}) \\ &\quad \text{since } A_n \setminus E_n = A_{n-1} \text{ and } \lambda \text{ is monotone} \\ &= \lambda(E_n) + \lambda(A_{n-1}) \\ &\quad \text{by what we have proved above} \\ &= \sum_{i=1}^{n} \lambda(E_i) \qquad \text{by induction.}\end{aligned}$$

Since λ is subtractive,

$$\begin{aligned}\lambda(A) &\geq \sum_{i=1}^{n} \lambda(A \cap E_i) + \lambda(A \setminus E) \\ &\geq \lambda(A \cap E) + \lambda(A \setminus E) \qquad \text{for arbitrarily large } n.\end{aligned}$$

Thus $E \in \mathcal{M}$, proving that $\mathcal{M}$ is a σ-field.

Now $\lambda(A \cap E) + \lambda(A \setminus E) \geq \lambda(A)$ by subadditivity. Hence

$$\lambda(A) = \sum_{i=1}^{\infty} \lambda(A \cap E_i) + \lambda(A \setminus E).$$

Replacing A by $A \cap E$ and noting that $\bigcup_{i=1}^{\infty} (A \cap E_i) = A \cap E$ gives the assertion.

(b) All properties are immediate except for countable subadditivity. If $A = \bigcup_{n=1}^{\infty} A_n$, we must show that $\mu^*(A) \le \sum_{n=1}^{\infty} \mu^*(A_n)$; and we may assume that $\mu^*(A_n) < \infty$ for all n. Given $\varepsilon > 0$, we may choose sets $E_{n,1}, E_{n,2}, \ldots \in \mathcal{E}$ with $A_n \subset \bigcup_k A_{nk}$ and $\sum_k \mu(A_{nk}) < \mu^*(A_n) + \varepsilon/2^n$. Then $A \subset \bigcup_{n,k} A_{nk}$ and $\sum_{n,k} \mu(A_{nk}) \le \sum_n \mu^*(A_n) + \varepsilon \le \sum_n \mu^*(A_n) + \varepsilon$. Since ε is arbitrary, $\mu^*(A) \le \sum_n \mu^*(A_n)$.

Now if $A \in \mathcal{E}$, then $\mu^*(A) = \mu(A)$ by restriction of μ; and if $A = \bigcup_n E_n$, $E_n \in \mathcal{E}$, then $\mu(A) \le \sum_n \mu(E_n)$ by (a) and (c). Thus μ is minimum over all such coverings of A by sets in $\mathcal{E}$, and $= \mu^*(A)$. Hence $\mu^* = \mu$ on $\mathcal{E}$.

(c) If $E \in \mathcal{E}$, $A \subset \Omega$, we must show that $\mu^*(A) \ge \mu^*(A \cap E) + \mu^*(A \cap E^c)$; we may assume $\mu^*(A) < \infty$. Given $\varepsilon > 0$, there are $E_1, E_2, \ldots \in \mathcal{E}$ with $A \subset \bigcup_n E_n$ and $\sum_n \mu(E_n) \le \mu^*(A) + \varepsilon$. Now

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) \le \mu^*\left(\bigcup_n (E_n \cap E)\right) + \mu^*\left(\bigcup_n (E_n \cap E^c)\right) \qquad \text{by monotonicity}$$

$$\le \sum_n \mu(E_n \cap E) + \sum_n \mu(E_n \cap E^c)$$

since μ^* is countably subadditive and $\mu^* = \mu$ on $\mathcal{E}$.

Similarly

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) \le \sum_n \mu(E_n \cap E) + \sum_n \mu(E_n \cap E^c).$$

Thus

$$\mu^*(A \cap E) + \mu^*(A \cap E^c) \le \sum_n \mu(E_n) \le \mu^*(A) + \varepsilon$$

and the result follows.

(d) If $A = B \cup N$ where $B \in \mathcal{F}_\mu$ and $N \subset M \in \mathcal{E}$ with $\mu^*(M) = 0$, then $B \in \mathcal{F}_{\mu^*}$ [since $\mathcal{E} \subset \mathcal{F}_{\mu^*}$ and $\mathcal{F}_{\mu^*}$ is a σ-field], so $\mu^*(B) = \mu(B)$. Also any set C with $\mu^*(C) = 0$ belongs to $\mathcal{F}_{\mu^*}$ by definition of μ^*. Countable additivity follows at once. Thus $\mathcal{F}_\mu \subset \mathcal{F}_{\mu^*}$. Therefore the completion of $\mathcal{E}(\mathcal{F})$ is included in $\mathcal{F}_{\mu^*}$.

Now assume μ is σ-finite on $\mathcal{F}_\mu$ and let $A \in \mathcal{F}_{\mu^*}$, $\varepsilon > 0$. By the argument of (c) there exist $A_1, A_2, \ldots \in \mathcal{E}$ with $\mu(A_n) < \infty$ and $\mu^*(A \cap A_n^c) < \varepsilon/2^n$. [Note that $\mu^*(A \cap A_n^c) = \mu^*(A \cap A_n) + \mu^*(A \cap A_n^c) - \mu^*(A \cap A_n)$.] However, if $A = \bigcup_n (A \cap A_n)$ [so that $A \cap A_n$ also belongs to $\mathcal{F}_{\mu^*}$], we have

$$\mu^*(A) \le \mu^*(A \cap A_n) + \mu^*(A \cap A_n^c) \le \mu^*(A \cap A_n) + \varepsilon/2^n$$

If $N = \bigcup_i N_i$, then $\bar{N} = \bigcup_i \bar{N}_i$, $A \subset \bar{A}$ and $\mu(\bar{A} - A) = \lambda(\bar{A} - A) = 0$. This argument applies to A, a union of a sequence of sets $A_i \in \mathcal{A}_i$ with $C_i \subset A_i$ and $\mu(A_i - C_i) < \varepsilon/2^i$. Let $C = \bigcup_i C_i$. Then $C \subset A$ and $\mu(A - C) = \mu(\bigcup_i (A_i - C_i)) < \varepsilon$. Thus $A \in \mathcal{A}_\mu$ and $\mu(A) = \mu(C) + \mu(A - C) < \mu(C) + \varepsilon$. This A belongs to the completion of $\mathcal{A}_\mathcal{R}$ relative to μ.

Section 12.4

1. Using the formulas of 12.43 the following results are obtained.

   $\mu_1 \times \mu_2(\{h\} \times [0,1]) = 0$, $\mu_1 \times \mu_2([0,1] \times \{y\}) = 0$

   $\lambda(\{y : \mu_1 \times \mu_2(\{h\} \times [0,y]) < \mu_1 \times \mu_2([0,1] \times [0,y])\}) = \lambda((0,1]) = 1$.

2. Let $C_n = \{x \in \mathbb{R}^1 : \ldots$ Use $x_i \le 1, i = 1, \ldots, n\}$. Then μ_n is finite on C_n. Given the Borel subsets B of C_n such that $\mu_n = m_n - \mu_n$ on $\mathcal{M}_n$ is monotone class including the class of finite disjoint unions of right semiclosed intervals in C_n (which all Borel subsets of C_n belong to the class from 12.2). If $B \in \mathcal{M}_n \cap C_n$, then $B \cap C_n \downarrow B$ hence $\nu + B \cap C_n \downarrow \nu + B$ and $\nu + B \cap C_n \in \mathcal{M}_n$ it follows that $\mu(\nu + B) = \mu(B)$.

   Now if $B \in \mathcal{M}(\mathbb{R}^1)$, then $B = A \cup N$, $A \in \mathcal{M}(\mathbb{R}^1)$, $C = D \in \mathcal{M}(\mathbb{R}^1)$ with $\mu(D) = 0$. Thus $\nu + B = \nu + A \cup N$, $\{\nu + A\} \cup \nu + D$, by Problem 5 $\nu + A \in \mathcal{M}(\mathbb{R}^1)$, $\nu + C \subset \nu + D \subset \mathcal{L}(\lambda^1)$. By what we have proved above $\mu(\nu + D) = \mu(D) = 0$ hence $\nu + N \in \mathcal{M}(\mathbb{R}^1)$ and $\mu(\nu + B) = \mu(B)$.

3. Let A be the unit cube $\{x \in \mathcal{X} : 0 < x_i \le 1, i = 1, \ldots, n\}$, and let $\varepsilon = p/q$. For any positive integer r we may divide each edge of A into rq equal parts so that A is decomposed into $(rq)^n$ subcubes $A_{ij}, \ldots, A_{(rq)^n}$ each with volume $1/(rq)^n$. By translation invariance, $\mu(A_i) = \mu(A_j)$ for all i, j, so if λ is Lebesgue measure we have

   $$\mu(A_i) = r^n \mu(A) = r^n \varepsilon = r^n p/q = r^n \lambda(A) q/p = \lambda(A)/q, \quad i = 1, \ldots, (rq)^n.$$

   Now any subinterval I of the unit cube can be expressed as the limit of an increasing sequence of sets B_n where each B_n is a finite disjoint union of subcubes of the above type. Thus $\mu = m$ on subintervals of the unit cube, and hence on all Borel subsets of the unit cube by the Caratheodory extension theorem. Since $\mathbb{R}^n$ is a countable disjoint union of cubes it is now clear that $\mu = m$ on $\mathcal{B}(\mathbb{R}^n)$.

4. (a) If $x - y = x_1 - x_2$, $x_1, x_2 \in A$, then x is equivalent to x_2 so that $x - y \in A - A$. Conversely, if $x - y \in A - A$ then x is connected to y by taking one member from each relation R_i. Thus $x - y$ is a rational number. Hence $x - y \in A$.

   (b) Let $x \in A$ be such that $x + y$ is the member of A that belongs to A. Then $x - y$ is a rational number, hence $x \in A$.

(b) If $0 \le r \le 1$, then $r = \sup\{x \in [0,2] : \text{dist}$

$$\sum_{n=1}^{\infty} \left| \mu_n - \lambda_n \right| \le r \le 1, \ldots, r \text{ rational} \}$$

$$= \mu\left(\bigcup_{n=1}^{\infty} [\lambda_n, \lambda_n + \mu_n) : 0 \le r \le 1, \ldots, r \text{ rational} \right) \quad \text{by (a)}$$

$$\le \mu[0,2] < \infty.$$

But $\mu[r, r+1] = \mu[0,1]$ by Problem 4; hence $\mu(r - \lambda) = \mu(r)$ and the $\mathcal{O}$ formula is false. Since $\mathcal{O}$ is a countable union of sets of μ-measure 0 by (a), $\mu(\mathbb{R}) = 0$, a contradiction.

b Let $f(x, y) = 1$ if $x = y \ge 0$, $f(x, y) = 0$ if $x = y < 0$ or if $x \ne y$. If $x = 0$, $y = 1$, $x = -1$, $y = -1$, $x = y$, then

$$A_{1,0} A_{0,1} f(x, y) = A_{1,0}(A_{0,1} f)(x, y) = A_{1,0} f(x, y + x)$$

$$= f(x + 1, y + x) = f(x, y)$$

$$= 1 \ne 0 = f(0, 1) = A_{0,1} f(0, 0) = A_{0,1} A_{1,0} f(x, y);$$

since F is not a constant function. Other examples: $F(x, y) = \max(x, y)$, $F(x, y) = x + y$, the largest integer less than or equal to $x + y$.

Section E.5

2. If $B \in \mathcal{B}(\mathbb{R})$,

$$\{\omega : m(\omega) \in B\} = \{\omega : \lambda(\cdot, \omega) \in B\} \cup \{\omega : \lambda(\cdot, \omega) \in A^c : M(\omega) \in B\}$$

$$= [\lambda(\cdot)^{-1}(B) \cap \Omega] \cup [\Lambda \cap \lambda^{-1}(B)];$$

which belongs to $\mathcal{F}$ since λ and χ are Borel measurable.

3. (a) Let f be a measurable function on Ω. $f = \sum_{n=1}^{\infty} c_n \chi_{E_n}$ where $c_n \ge 0$, $E_n \in \mathcal{B}^k$; for all $\lambda > 0$, there exist $x_1, \ldots, x_n \in \mathbb{R}^k$ such that $|x_i - x_j| \ge \lambda$ and $x_i \in E_n$, or $f(x_i) = f(x_j) = 1/n$. We show that the E_n are closed. Let x_m be a sequence of points in E_n with $x_m \to x$. Let $\delta > 0$ and $N = \{x : |x - x_m| < \delta\}$. Then $x_m \in N$ for large x_m and since $x_m \in E_n$, there are points x_{i_m} such that $x_{i_m} \in N$ such that $|f(x_{i_m}) - f(x_m)| \ge 1/n$. Then $|x_{i_m} - x_m| < \delta$ and $|x_m - x| < \delta$, so $|x_{i_m} - x| < 2\delta$. So that $x \in E_n$.

The result is true for a function from an ordinary topological space to a metric space $(\mathcal{X}, d)$. Take $\Omega_n = \{x \in \mathcal{X}$ for every neighborhood $\mathcal{N}$ of x, there exist $x_1, x_2 \in \mathcal{N}$ such that $d(f(x_1), f(x_2)) \ge 1/n\}$. The above proof goes through with "sequence" replaced by "net."

The result is false if no assumptions are made about the topology of the range space. For example, let $\Omega = [0, 2]$, with open sets

$\mathbb{R}, \mathcal{B}, m$) and $[0,1]$. Define $f: \Omega \to \mathbb{R}$ by $f(\omega) = \omega(0) - \omega(1)$. Then the set of discontinuities is $\{t : \omega(t) \neq \omega(t-)\}$, which is not an F_t.

(b) This follows from part (a) because the continuous f cannot be approximated on a countable union of closed sets C_n. If this were possible then each C_n would have empty interior since every nonempty open set contains rational points. But then $\bigcup C_n$ is of category 1 in $\mathbb{R}$. It follows that A is of category 1 in itself, contradicting the Baire category theorem.

6. By Problem 11 at the end of Section 3.3, each $\sigma(Y)$-measurable Z is the limit of a sequence of simple functions. Hence there are only c simple functions on $\sigma(Y)$. Since a Borel measurable function is the limit of a sequence of simple functions, there are $c^{\aleph_0} = c$ Borel measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$. By 1.3.8 there are only c Borel measurable functions from $\mathbb{R}^n$ to $\mathbb{R}^n$.

7. (a) Since the F_n are measures, $\sum_{n=1}^{\infty} F_n(A_n) = F_n(\Omega) = 1$, and it follows quickly that the a_n satisfy the hypotheses of Scheffé's Lemma. If $\{n_k\}$ is the sequence given by the lemma, let $A = \{\omega : \omega_{n_k} = 1 \text{ for infinitely many } k\}$. Then

$$\left| P_n(B) - \sum_{\omega \in B} P(\{\omega\}) \right| = \left| P_n(B) - \sum_{\omega \in B \cap A} P(\{\omega\}) \right| \leq P_n(B) - \sum_{\omega \in B \cap A} P(\{\omega\}),$$

and it follows that P_n converges to the contradiction. Thus P is not child-y measure: $P(A_1) = \sum_{\omega \in A_1} P(\{\omega\}) = 0$ then $P(\{\omega\}) = 0$ for some $\omega \in A_1$. Thus $P_n(A_n) < \varepsilon$ for large n, say for $n \geq n_0$. Since the A_n decrease, we have $\sup_{\omega \in A_n} P(\{\omega\}) \leq \varepsilon$ for $k \geq k_0$ and since $A_n \downarrow A$, the ω in A_1 satisfy $P(\{\omega\}) = 0$ for $n = 1, 2, \ldots, n_{k_0}$.

$P_n(A_n) < \varepsilon$ for $k > k_0$. Thus $\sup_{\omega \in A_1} P(\{\omega\}) = 0$ for some A_1 in $\mathcal{F}$.

(b) Without loss of generality, assume $P(\{\omega\}) > 0$ for all ω. Take Ω to be the countable set $\{\omega_1, \omega_2, \ldots\}$ and set $P_n(\{\omega_n\}) = 1 - P_n(\{\omega_n\}) = P(\{\omega_n\})$. The P_n are now probability measures and the result follows from part (a).

Section 3.6

2. $\int \sum_{n=1}^{\infty} |f_n| \, d\mu = \sum_{n=1}^{\infty} \int_{\Omega} |f_n| \, d\mu < \infty$, because $\sum_{n=1}^{\infty} |f_n|$ is integrable and therefore finite a.e. Thus $\sum_{n=1}^{\infty} f_n$ converges a.e. to a finite valued function.

Let $g_n = \sum_{k=1}^{n} f_k$. Then $g_n \to \sum_{k=1}^{\infty} f_k$ an integrable function. By the dominated convergence theorem, $\int_{\Omega} g_n \, d\mu \to \int_{\Omega} g \, d\mu$, that is

$$\sum_{k=1}^{\infty} \int_{\Omega} f_k \, d\mu = \int_{\Omega} \sum_{k=1}^{\infty} f_k \, d\mu.$$

3. Let $q = (r, s)$ and let $a = x_0 < x_1 < \cdots < x_m = b$. Then

$$\begin{aligned} f(r, s) - f(q) &= \int_a^b f(x, s)\,dx - \int_a^b f(x, r)\,dx \\ &= \int_a^b \left[\frac{f(x, s) - f(x, y)}{s - r} - \frac{f(x, r) - f(x, y)}{r - r} \right] dy \end{aligned}$$

By the mean value theorem,

$$\frac{f(x_i, s) - f(x_i, y)}{s - y} - f_2(x_i, t_i) \quad \text{and} \quad \frac{f(x_i, r) - f(x_i, y)}{r - y} = f_2(x_i, t_i)$$

for some $t_i = t_i(y)$ between r and s. By hypothesis, $f_2(x, t)$ is bounded on R and f_2 is integrable, and the result now follows from the dominated convergence theorem since $|f(x_i, s) - f(x_i, y)| / |s - y| \le M$ and $f(x, t)$ is continuous as a function of t.

8. Let μ be Lebesgue measure on I. If f is an indicator, $f = \chi_E$, $E \subset I$, we can take the partition $\{x_0, x_1, x_2\}$ with $x_0 < \inf E < \sup E < x_2$, which follows by monotonicity of μ as in Problem 4 in 7.4.6. The passage to simple functions, nonnegative measurable functions, and arbitrary measurable functions is done as in 7.4.12.

Section 7.7

2. (a) If $\int$ is Riemann–Stieltjes integrable, $\int = \int_a^b f\,d\alpha$, then by Section 7.7.1(a) The theorem of Lebesgue on the set of discontinuities of f is critical. The set of measure 0 appears as the exceptional set of measure 0 in the definition of Riemann integrability. Given $\varepsilon > 0$, if f is continuous at x, then α is continuous at x [(a)]. Conversely, if α is discontinuous at x, then $\alpha(x) < \alpha(x+) - \varepsilon < \alpha(x+)$ for some $\varepsilon > 0$. The result that f is continuous at x implies $\alpha(x) = \alpha(x-) = \alpha(x+)$ is true even if x is an endpoint. As in 7.7.1(a), f is Riemann–Stieltjes integrable.

   (b) This is done essentially as in 7.7.1(b).

5. (a) By definition of the improper Riemann integral, f must be Riemann integrable (hence continuous a.e.) on each finite interval, and the result is true. For the counterexample, let $f(x) = (-1)^n / n$ for $n \le x < n + 1$, n a positive integer, $f(x) = -1$ for $x < 1$, f is an odd function. Then the limit $\alpha(x)$ does not exist. Indeed, any value $\alpha(x)$ between $\alpha(y)$ and $\alpha(y+)$ does not exist. Also $\alpha(x) \to -\infty$ as $x \to -\infty$, $\alpha(x) \to \infty$ as $x \to \infty$.

   (b) Define

$$f_n(x) = \begin{cases} f(x) & \text{if } -n \le x \le n, \\ 0 & \text{elsewhere.} \end{cases}$$

Then $f_n \to f$. Hence f is measurable relative to the completed σ-field, and $\int f_n \, d\mu \to \int f \, d\mu$ in the monotone convergence theorem. But $\int f_n \, d\mu = \int_{[0,1]} f_n \, d\lambda$ by 1.7.(b) and $\int_{[0,1]} f_n \, d\lambda \to \int_{[0,1]} f \, d\lambda$ by hypothesis; the result follows.

For the second example, take

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{x}}, & 0 < x < 1 \quad (n = 1), \quad \dfrac{1}{n-1} < x < \dfrac{1}{n} \quad (n = 3, 4, \ldots) \\ 0, & x = 0 \end{cases}$$

We have $\int_0^1 f \, d\lambda = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$, but

$$\int_0^1 |f| \, d\lambda = 1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty,$$

so that f is not Lebesgue integrable on $[0, 1]$.

Chapter 2

Section 2.7

1. Let $\Omega = \{\omega_1, \omega_2, \ldots\}$, $P(\{\omega_n\}) = p_n > 0$, let $\mathcal{A}$ be the σ-field $\mathcal{P}(\Omega)$, and let $\mathcal{B}$ be the σ-field $\{\emptyset, \Omega\}$. Then $\mathcal{B} \subset \mathcal{A}$, and for any $X \in \mathcal{A}$,

$$X^{\mathcal{B}}(\omega) = E(X \mid \mathcal{B})(\omega) = \int_\Omega X \, dP = \sum_{n=1}^{\infty} X(\omega_n) p_n$$

since $X^{\mathcal{B}} = X$ on Ω and $X^{\mathcal{B}} = 0$ on $\emptyset$. Similarly,

$$X^{\mathcal{A}}(\omega_n) = E(X \mid \mathcal{A})(\omega_n) = \frac{1}{p_n} \int_{\{\omega_n\}} X \, dP$$

since $X^{\mathcal{A}} = X$ on Ω and $X^{\mathcal{A}} = 0$ on $\emptyset$. The result follows.

4. If $E_1, \ldots, E_n$ are disjoint sets in $\mathcal{A}$ with all $E_i \in A_0$,

$$\sum_{i=1}^{n} |\lambda(E_i)| = \sum_{i=1}^{n} |\lambda^+(E_i) - \lambda^-(E_i)| \le \sum_{i=1}^{n} [\lambda^+(E_i) + \lambda^-(E_i)]$$

$$= |\lambda| \left(\bigcup_{i=1}^{n} E_i \right) \le |\lambda|(\Omega).$$

Thus the sup of the terms $\sum_{i=1}^{n} |\lambda(E_i)|$ is at most $|\lambda|(\Omega)$. But

$$|\lambda|(\Omega) = \lambda^+(\Omega) + \lambda^-(\Omega)$$

$$= \lambda(\Omega) + 2\lambda^-(\Omega)$$

$$= |\mu(A \cap D)| - |\mu(B \cap D)|$$

since $\nu(E \cap D) \geq 0$ and $\nu(E \cap D) \geq 0$

$$|\mu(F)| + \nu(F)$$

and the result follows.

Section 4.3

3. Let $A_n = \{x : |g(x)| > n^{-1}\}$, $n = 1, 2, \ldots$, and let $A = \bigcup_{n=1}^{\infty} A_n$. Now

$$\infty > \int_A |g|\,d\mu \geq \frac{1}{n}\mu(A_n).$$

Hence $\mu(A_n) < \infty$. For the example, let g be Lebesgue measure on $(\mathbb{R}^1)$ and let $\mu(g)$ be any strictly positive integrable function, such as $g(x) = e^{-|x|}$. In this case $A = \mathbb{R}$, so that $\mu(A) = \infty$.

4. If f is an indicator f_D, the result is true by hypothesis. If f is a nonnegative simple function $\sum_{i=1}^{n} a_i f_{D_i}$ for $D_1, \ldots, D_n$ disjoint sets in $\mathcal{S}$, then

$$\int_B f\,d\lambda = \sum_{i=1}^{n} a_i \lambda(B \cap D_i) = \sum_{i=1}^{n} a_i \int_{D_i} g\,d\mu = \sum_{i=1}^{n} a_i \int_B f_{D_i} g\,d\mu$$

$$= \int_B fg\,d\mu \qquad \text{by the additivity theorem.}$$

If f is a nonnegative Borel measurable function, let $f_1, f_2, \ldots$ be nonnegative simple functions increasing to f. By what we have just proved, $\int_B f_n\,d\lambda = \int_B f_n g\,d\mu$, hence $\int_B f\,d\lambda = \int_B fg\,d\mu$ by the monotone convergence theorem. Finally, if f is an arbitrary Borel measurable function, write $f = f^+ - f^-$. By what we have just proved,

$$\int_B f^+\,d\lambda = \int_B f^+ g\,d\mu \qquad \int_B f^-\,d\lambda = \int_B f^- g\,d\mu$$

and the result follows from the additivity theorem.

6. (a) In the definition of $\int f\,d\mu$ we may assume without loss of generality that the F_j are disjoint. If $\{A_i\}$ is a disjoint union of the $\{F_j \cap E\}$, then

$$\sum_{j=1}^{n} \mu(E \cap F_j) = \sum_{j=1}^{n} \sum_{i} \mu(A_i \cap F_j) \leq \sum_{i=1}^{m} \sum_{j=1}^{n} \mu(A_i \cap F_j)$$

$$= \sum_{i=1}^{m} \sum_{j \in I} \mu(A_i \cap F_j) = \sum_{i=1}^{m} \mu(A_i),$$

Thus $\lambda([a,c)\cap A)\ge\sum_n p(a_n)$. Now to show the reverse inequality, we may assume $\lambda([a,b)\cap A)<\infty$; hence $p([a,b))<\infty$. For each k, there is a partition $\{E_1,\dots,E_{j_k}\}$ of A such that

$$\sum_{j=1}^{j_k}[\sup E_j-\inf E_j]<p([a,b))-\frac{1}{k}\qquad\text{(proved)}.$$

Then for any $\varepsilon>0$,

$$p([a,b))<\sum_{j=1}^{j_k}[\sup E_j-\inf E_j]+\frac{1}{k}\le\sum_{n=1}^{\infty}p([a_n,c_n))+\varepsilon.$$

Since ε and k are arbitrary, the result follows.

(b) If $E_1,\dots,E_n$ are disjoint measurable subsets of A,

$$\sum_{j=1}^{n}[\lambda(A\cap E_j)]=\sum_{j=1}^{n}\lambda(A\cap E_j)=\sum_{j=1}^{n}p(E_j)\le\lambda([a,b)\cap A),$$

proving $[m]+\lambda(A)\le[p]-m(a)=m(b)-m(a)=p([a,b))$. Inequality follows from the definition of outer variation.

(c) If $m(a,)=0$ and $\lambda(A,m)=\lambda(-\infty,b)$, then $m(a_n,b)\downarrow m(a,b)$ and $\lambda(A,[a_n,b))\uparrow\lambda(A,[a,b))$, so $[m_{a_n}]\to[m_a]$ by Exercise 4.2.5(a). Hence $[m_a]=m(b)-m(a)=0$.

(d) This has been established when A is any open set (see 4.2.5), so assume A is compact, say $A=A_1\cap G_2$. If $p(A)=0$, then $\lambda(A,[a,b))=\lambda(G_2,A_1)=0$. Hence λ vanishes on all sets $A_1\cap G_2$ and G_2 open. By 2.2.5(f) $\lambda(B,[a,b))=0$ for $B\in\mathcal{B}$ hence by (b), m_a is p. The converse is clear since $p([a,b))=\lambda(A,[a,b))$.

(e) The proof is the same as in 2.2.5(d).

(f) See 2.2.5(d).

(g) The "if" part is immediate in $\mathbb{R}^1$ since for the Young–Stieltjes integral $\int_a^b f\,dm_a=m(b)-m(a)$. Since $\lambda(A,[a,b))=0$ by (c), $\int_A f\,dm_a=0$ by 2.2.5(e) because $\lambda(A,\cdot)=0$.

Section 2.3

2. We have

$$\int_a^b f'(x)\,dx=\int_a^b\left[\lim_{\lambda\to 0}\frac{f(x+\lambda)-f(x)}{\lambda}\right]dx$$

where we may assume $\lambda>0$.

$$\le\liminf_{\lambda\to 0}\int_a^b\left[\frac{f(x+\lambda)-f(x)}{\lambda}\right]dx$$

$$= \lim_{n\to\infty} \left| \int_a^{x_n} f(t)\,dt - \int_a^{x} f(t)\,dt \right|$$

$$\text{Define } g(t) = f(t), \quad a \le t \le b, \qquad g(t) = f(x), \quad x \le t \le x_n$$

$$\le \lim_{n\to\infty} \int_x^{x_n} |f(t) - f(x)|\,dt \qquad \text{since } f \text{ is continuous at } x$$

$$= 0.$$

Alternatively, let μ be the Lebesgue-Stieltjes measure corresponding to λ (discussed in Problem 5.2.3), and let m be Lebesgue measure. Write $u = \int_a^x f\,d\lambda$ where $u(x) = 0$ and $u(b) = c$. By 2.8.5 and 2.8.6,

$$\int_a^b f(x)\,d\lambda = \int_a^b f\,d\mu$$

$$= \int_a^b g\,d\mu$$

$$= \int_a^b \frac{dg}{dm}\,dm = u_1[a,b] \le g(b) - g(a) = \int_a^b g\,dx.$$

6 (a) Since A is positive and symmetric, $\langle x, y\rangle = \langle Ax, y\rangle$ defines an inner product on $\mathbb{R}^n$ by Problem 5.1.14. $A = \alpha I$ for some constant $\alpha > 0$. Now if A_1 and A_2 are linear transformations on $\mathbb{R}^n$ then

$$\min\{\alpha_1, \alpha_2\}\|x\|^2 = \min\{\alpha_1\|x\|^2, \alpha_2\|x\|^2\} \le \langle A_1 x, x\rangle + \langle A_2 x, x\rangle$$

and

$$\langle (A_1 + A_2)x, x\rangle = \langle A_1 x, x\rangle + \langle A_2 x, x\rangle$$

where

$$\langle A_j x, x\rangle = \alpha_j \|x\|^2.$$

Since $\det(\alpha_j I) = \det A_j$, it suffices to assume that A is an orthogonal transformation. Now if Q is an orthogonal matrix $[x \to Qx]$, $\det Q = \pm 1$ since $\det(Q) = \det(Q^T)$. If $e_1, \ldots, e_n$ is the standard basis for $\mathbb{R}^n$, then A fulfills conditions of the following three categories:

(1) $Ae_1 = -e_1, \ldots, Ae_n = -e_n$, where n is even. Then $\det A = 1 = \det A$.

(2) $Ae_1 = -e_1, \ldots, Ae_n = -e_n$, where n is odd. Then $\det(A) = -1 = \det A$.

(d) $Ax_j = e_j$ $(1 \le j \le n)$. Also $\mathbf{x} = \sum_{j=1}^{n} x_j e_j$. Therefore $\mathbf{x} = 0$ corresponds to the following system of linear equations: $x_j = 0$ $(1 \le j \le n)$. We must assume $i = 1, j = 2$. Then

$$A \cap Q = \left\{ x \sum_{i=1}^{n} a_{ij} x_j : 0 < x < 1 \quad (i = 1, \ldots, n) \right\}$$

$$= \left\{ y = \sum_{j=1}^{n} a_{ij} x_j : 0 < x_j < \quad (j \neq 1), \quad 0 < x < b - \right\}$$

If $A_1 = \{y \in A(Q) : y_1 < 1\}$, $A_2 = \{y \in A(Q) : y_1 > \}$, and $A_3 = \{y \in A(Q) : y_1 = 1\}$, then $A_1 \cap A_2 = \emptyset$; for if $y \in A_1$, then $y \in A_2$ and if $y = \sum a_j x_j \in A_3 \cap A(Q)$, then $x_1 = 1 \le b$. Therefore,

$$m(A) = m(A(Q)) = m(A_1 \cup A_2) = m(A_1) + m(A_2) = m(A_1) + \quad = m(A_1 \cup A_2 \cup A_3) = m(Q) = 1.$$

(b) Let $x \in V$ and define $S(x) = \{(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \in C\}$. Then if C is a cube containing Q, we have

$$T(x) \cap C = \{T(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \in C\} = T(x) + S(C)$$

$$= T(x) + S(Q).$$

Therefore, $m(T(x) \cap C) = m(S(Q))$, so that the invertibility of T at x is equivalent to differentiability of S at Q, and the derivatives, if they exist, are equal. Now the function $x \to S$ is linear with the Jacobian matrix of T at x and the result follows.

(c) Let $A = A(Q)$, and assume $S(y) = A^{-1}(T(y))$, $x \in V$. The Jacobian matrix of S at 0 is the identity matrix, and $S(0) = A^{-1}(0) = 0$. If we can show that $m(S(E)) = m(E)$, $E \in \mathcal{M}(\mathbb{R}^n)$, $E \subset$ Euclidean at 0 with derivative I, then

$$\left| \frac{m(S(C))}{m(C)} - 1 \right| < \frac{\varepsilon}{\det A}$$

for sufficiently small cubes C containing 0. Then by $m(T(C)) = m(S(C)) = \det A \cdot m(S(C))$, hence

$$\frac{m(T(C))}{m(C)} = |\det A| \cdot \frac{m(S(C))}{m(C)} = |\det A| \left| \frac{m(S(C))}{m(C)} - 1 \right| < |\det A| \cdot \varepsilon$$

so that T is differentiable at 0 with derivative $\det A$.

(d) (i) If $x \in C_1$, then $|x| \le \sqrt{2}/3 < 1/2$. Hence $T(x) = x/3 \in C_0 = J(0; 1/2)$. Therefore $T(C_1) \subset C_0$.

(ii) If $x \in C_2$, then $|T(x) - x| \ge \alpha/3$ as above, and $\alpha/3 = J(0) - 1/2$. Therefore $T(x) \notin C_1$.

(iii) We have $T(x_j + \varepsilon) = x_j + \varepsilon/3$, and

$$\alpha/3 = \frac{\alpha/3}{1 - \alpha/3} < \frac{1}{2^{j+1}} \cdot \frac{1}{3^j} = \frac{1}{2^{j+1}3^j},$$

and the result follows.

(iv) If $x \in C_{n+1} = T(C_n)$, then $x = T(y)$ for some $y \in T(C_n)$, and $y \in C_n$, so $J(x) \subset J(y)$. Hence $J(x) \subset J(C_n)$. Since $J(C_n)$ is the disjoint union of the sets $J(C_j) \setminus T(J(C_j))$ and $C_n = J(C_n) \setminus J(C_n)$, the first set is open since T is an open map, and the second set is open by (iii). By (iii), $C_n \cap T(J(C_n)) = \emptyset$, so by connectedness of $J(C_n)$ we have $J(C_n) \setminus T(J(C_n)) \subset J(C_n)$, that is, $C_n \subset T(J(C_n))$. So $T(C_n) \subset C_n$ by (i).

Therefore $m(J(C_n)) < m(T(J(C_n))) + m(C_n)$, that is,

$$1 - 2\alpha/3^{n+1} < (1 - \alpha/3)(1 - 2\alpha/3^n) + 1 - 2\alpha/3^n.$$

Thus $1 - \alpha < (1 - \alpha/3)(1 - 2\alpha/3^n) + 1 - \alpha/3$, and the result is derived.

(v) Assume $m(E) = 0$ and $J(E) = C$. By (iv), (1), and the fact that

$$E \subset \bigcup_{n=1}^{\infty} E \cap \left[\sup_{|x-y|<\delta} \frac{|J(x) - J(y)|}{|x - y|} < n \right],$$

we can find a compact set K and positive integers n and j such that $m(K) > 0$, $J(K) > 0$, and $|J(x) - J(y)| < n|x - y|$ for all open cubes Q containing a point of K and having diameter less than $1/j$. Let $\varepsilon > 0$, choose an open set $O \supset K$ with $m(O) < m(K) + \varepsilon$.

Now partition $\mathbb{R}^d$ into disjoint (partially closed) cubes S of diameter less than $1/j$ and small enough so that $S \cap K \neq \emptyset$ implies $S \subset O$. Let us indicate that since K and $\mathbb{R}^d \setminus K$ are both measurable, we may find open cubes $Q_1 \supset K$, $Q_2 \supset \mathbb{R}^d \setminus K$, such that $m(Q_1 \cap Q_2) < \varepsilon m(Q_1)$ and $\operatorname{diam} Q_1 < 1/j$. Then

$$J(K) \le \sum_{i=1}^{\infty} J(Q_i) \le \sum_{i=1}^{\infty} n|Q_i| < n \sum_{i=1}^{\infty} m(Q_i) < 2n \sum_{i=1}^{\infty} m(S_i)$$
$$< 2n(m(O) + 2\varepsilon).$$

Since ε is arbitrary, $J(K) = 0$, a contradiction.

(c) If f is an indicator χ_E, let $E = T^{-1}(a,b)$, so that $B = Y(E)$. Then

$$\int_X g\,\chi_E\,d\mu = m(Y(E)) \quad \text{and} \quad \int_0^1 \chi_E(Y(x))\,dx = \int_E |Y'(x)|\,dx,$$

and the result follows from part (b). The usual passage to simple functions, nonnegative measurable functions and finally to measurable functions completes the proof.

Section 2.4

1. (a) If $\Gamma = \{x_1, \ldots, x_n, 0, \infty, \ldots\}$, then $\int_\Gamma f\,d\psi = \sum_{j=1}^n a_j$ by definition of the integral of a simple function. If $\Gamma = \{x_1, x_2, \ldots\}$ with all $x_j \neq 0, \infty$, $\int_\Gamma f\,d\psi = \sum_{j=1}^\infty a_j$ by the result for simple functions and monotone convergence theorem. If the x_j are not nonzero, then $\int_\Gamma f\,d\psi = \int_{\Gamma'} f\,d\psi = \int_{\Gamma''} f\,d\psi = \sum_{x_j \neq 0} a_j = \sum_{x_j \neq \infty} a_j$ as also in the text. Finally, if the x_j are complex,

$$\int_\Gamma f\,d\psi = \sum_{x_j \neq \infty} \operatorname{Re} a_j + i \sum_{x_j \neq \infty} \operatorname{Im} a_j$$

provided $\operatorname{Re} f$ and $\operatorname{Im} f$ are integrable; since $|\operatorname{Re} a_j|, |\operatorname{Im} a_j| \leq |a_j| \leq |\operatorname{Re} a_j| + |\operatorname{Im} a_j|$, this is equivalent to $\sum_{x_j \neq \infty} |a_j| < \infty$.

(b) If $\Gamma \neq \Gamma'$ except for a set E of μ-measure 0, then $\int_\Gamma f\,d\psi = \int_E f\,d\mu$ by definition of the integral of a simple function. If $f \geq 0$, then $\int_\Gamma f\,d\psi \geq \int_{\Gamma'} f\,d\psi$ for all finite $\Gamma' \subset \Gamma$; hence $\int_\Gamma f\,d\psi \geq \int_E f\,d\mu$ for uncountably many α, then $\sum_\alpha \mu(E_\alpha) = \infty$; hence $\int_E f\,d\mu = \infty$ also. If $\mu(E_\alpha) > 0$ for only countably many α, then $\int_E f\,d\mu = \sum_\alpha \int_{E_\alpha} f\,d\mu$ by the monotone convergence theorem. The remainder of the proof is as in part (a).

8. Apply Hölder's inequality with $g = 1$ and f replaced by $|f|^p$ and p replaced by $q = p/(p-1)$ to obtain

$$\int_{\mathbb{R}^n} |f|^p\,dx \leq \left(\int_{\mathbb{R}^n} |f|^{p+\varepsilon}\,dx\right)^{p/(p+\varepsilon)} \left(\int_{\mathbb{R}^n} 1\,dx\right)^{\varepsilon/(p+\varepsilon)}.$$

Therefore $\|f\|_p \leq \|f\|_{p+\varepsilon}\,(\mu(\mathbb{R}^n))^{\varepsilon/p(p+\varepsilon)}$, as desired.

9. We have $\int_E |f|^p\,dx \leq \|f\|_\infty^p \mu(E) < \infty$ so $f \in L^p(E)$ and $\|f\|_{L^p(E)} \leq \|f\|_\infty$. Now $\|f\|_\infty = \inf\{\alpha : |f(x)| \leq \alpha \text{ a.e.}\} = \inf\{\alpha : \mu(\{x : |f(x)| > \alpha\}) = 0\}$. Thus

$$\int_{\mathbb{R}^n} |f|^p\,dx = \int_E |f|^p\,dx \geq \int_E \|f\|_\infty^p\,dx = \|f\|_\infty^p \mu(E).$$

Since $\mu(A) > 0$ by definition of $\|f\|_\infty$ we have $\lim_{n\to\infty} \|f\|_n \geq \|f\|_\infty$. If $\|f\|_\infty = \infty$, let $A_M = \{x : |f(x)| \geq M\}$ and show that

$$\liminf_{n\to\infty} \|f\|_n \geq M;$$

since M can be arbitrarily large the result follows.

If $\mu(\Omega) = \infty$ or $\geq$ still fails, but $\lim_{n\to\infty} \|f\|_n = \|f\|_\infty$ if $\|f\|_\infty = \infty$ in the above argument, then $\|f\|_n = \infty$ for all n. However, if μ is Lebesgue measure on $\mathcal{M}(\mathbb{R})$ and $f(x) = 1$ for $n \leq x \leq n+1$ ($n = 0, \pm 1, \pm 2, \ldots$) and $f(x) = 0$ elsewhere, then $\|f\|_n = \infty$ for $n = 0, \pm 1, \pm 2, \ldots$ but $\|f\|_\infty = 1$.

11. (a) We have $|\int_E f d\mu| \leq \int_E |f| d\mu$. But $\omega \in E$ implies $f(\omega) \neq 0$ hence $|f(\omega)| > 0$ and thus $\int_E |f| d\mu > 0$ since $m(E) > 0$ if $\mu(\Omega) > 0$. Thus

$$\frac{1}{\mu(\Omega)} \int_E f \, d\mu = 0 \left| = \frac{1}{\mu(\Omega)} \int_E |f| \, d\mu > 0 \right|,$$

hence $\int_E f(\omega) P(d\omega) = 0 \geq 0$, a contradiction. Therefore $P(E) = 0$ that is $P(\omega : f(\omega) = 0) = 0$. Since $f = f(\omega)$ is a generic value of the r.v. $f = f(\omega)$, the result follows.

(b) Let $\mu = |\lambda_1| + \cdots + |\lambda_n|$ if $E_1, \ldots, E_n$ are disjoint measurable subsets of Ω,

$$\sum_{j=1}^{n} |\lambda(E_j)| = \sum_{j=1}^{n} \left| \int_{E_j} h \, d\mu \right| \leq \sum_{j=1}^{n} \int_{E_j} |h| \, d\mu$$

$$\leq 1 \cdot \sum_{j=1}^{n} \mu(E_j) \leq \mu(\Omega) < \infty.$$

Thus $\mu(\Omega) < \infty$ and since $0 < \mu < 1$, we have $\mu(\Omega) = 0$. If $A = \{\omega : |h(\omega)| < 1\}$, $\Omega = \{|h| < 1\} \cup \{|h| \geq 1\}$ (a.e.), then $\mu(A) = 0$ so that $|h| \geq 1$ a.e. $|\mu|$.

Now if $\mu(E) > 0$, then $|\lambda(E)| = |\int_E h d\mu| < \mu(E) \|h\|_\infty (= 1)$ where $\|h\|_\infty = \inf\{c : |h| < c$ a.e. $|\mu|\}$. By (a), $h(\omega) = e^{i\theta}$ a.e. $|\mu|$ for some $\theta \in [0, 2\pi)$ a.e. $|\mu|$.

12. If $\Sigma = \sigma(\mathcal{C})$, $\int_\Omega f d\lambda = \int_\Omega h d|\lambda| = \lambda(E)$; let $f(\omega) = e^{i\theta}$. Then $\int_\Omega f h d\mu = \int_\Omega f d\lambda = \lambda(E) = \int_\Omega \chi_E d\lambda$ by definition of λ. It follows consistently that $\int_\Omega f^n h d|\mu| = \int_\Omega \chi_E d|\lambda|$ readily extends to f as a complex valued simple function. If f is a bounded, complex valued Borel measurable function, by 1.5.5(b) there are simple functions $f_n \to f$ with $|f_n| \leq |f|$. By the dominated convergence theorem, $\int_\Omega f h d\mu = \int_\Omega f d\lambda$. If $f = \overline{h}/|h|$, we obtain $|\mu|(E) = \int_E h d|\mu|$.

(iii) $\int_0^1 f(x)\,x^n\,dx = 0$ for all n. Hence by the Lemma $f(x) = 0$ a.e. But if $g(x) = \exp(1/x^2)$ and $h(x) = e^{1/x^2}$, then $g'(x) = -2x^{-3}g(x)$, $h'(x) = -2x^{-3}h(x)$, so $gh' = hg'$ on $(0,1)$, and $\int_0^1 (gh' - hg')\,dx = g(1) - h(1) = e - e = 0$.

12. If f_1, f_2 are two absolutely continuous functions, let $G = f_1 - f_2$ and let C be the constant value of each side.

$$\int_0^x G(t)\,dt = Cx \quad (0 \le x \le 1).$$

Hence $\int_0^x G(t)\,dt - \int_0^y G(t)\,dt = C(x-y)$. For every $\varepsilon > 0$, there are pairwise disjoint intervals $(a_1, b_1), \dots, (a_n, b_n)$ such that $\sum (b_i - a_i) < \varepsilon$ and $\sum |G(b_i) - G(a_i)| > 1$. (Choose greedily: let $\sup_{a_1 < x < b_1} |G(x) - G(a_1)| > 1$, and so on; since $\lim_{x \to a_1} |G(x) - G(a_1)| = 0$, countably many such intervals exist.) Then $|C| > 1$, contradicting $C = 0$. Hence $G = 0$ a.e., so $f_1 = f_2$ a.e.

Section 2.5

4. Let $E_n = \{x : |f(x)| \le n\}$, $E_0 = \{x : f(x) = \infty\}$, $\{E_n\}_{n=0}^{\infty}$. Then

$$\int_E g\,d\mu = \sum_{n=0}^{\infty} \int_{E_n} g\,d\mu,$$

and the proof proceeds just as in 2.5.4.

5. Let $\{f_n\}$ be a bounded sequence converging to f pointwise a.e. on E by Theorem 1. By Egoroff's theorem, it is uniformly convergent on E except on a set of arbitrarily small measure. Now $\int_E f_n\,d\mu - \int_E f\,d\mu = \int_E (f_n - f)\,d\mu$; then on the set where $|f_n - f| < \varepsilon$ we have $\left|\int_E f_n\,d\mu - \int_E f\,d\mu\right| < \varepsilon\,\mu(E)$ for all sufficiently large n. (This last argument shows that every subsequence $\{f_{n_k}\}$ of $\{f_n\}$ converges to f a.e., so that $\int_E f_n\,d\mu \to \int_E f\,d\mu$ by 2.6.9, a convergence-in-measure result.)

Section 2.6

4. By Fubini's theorem,

$$\mu(E) = \int_{\mathbb{R}^2} \chi_E\,d\lambda = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \chi_E(x,y)\,dx\right] dy = \int_{\mathbb{R}} \mu(E(y))\,dy = 0.$$

Similarly, $\mu(E^x) = 0$ for all x. The result follows since $f \ge 0$ a.e. $[\mu]$ and $\int_E f = 0$ implies $f = 0$ a.e.

7. (a) Let

$$A_{kn} = \left\{x \in \Omega : \frac{k}{n} \le f(x) < \frac{k+1}{n}\right\},$$

$$B_{kn} = \left\{(x,t) \in \Omega \times \mathbb{R} : \frac{k}{n} \le t < \frac{k+1}{n}\right\}.$$

$x_k = 1, 2, \ldots, n$ $k = 1, 2, \ldots, n$, when $k = n$, include the right endpoint as well. Then

$$\Omega = \prod_{n=1}^{\infty} \bigcup_{k=1}^{n} (A_{nk} \times B_{nk}) \in \mathscr{R}.$$

If f is any real-valued function on a subset of Ω, replace 0^0 by the constant 1 in the definition of x_n.

(b) Assume $B \in \mathscr{B}$. Each $x \in \Omega$ is completely determined by the set $\{y \in B : y \text{ is the first coordinate of only countably many points } (y, y_2), y_2 \in \mathbb{R}\}$ such that $x_n = x_n$ for $n \geq 1$. (If there are only finitely many points $(y, y_2), \ldots, (y, y_m)$, take $x_n = x_n$ for $n \geq m$.)

Thus $B = \bigcup_{n=1}^{\infty} G_n$, where G_n is the graph of the function f_n defined by $f_n(x) = x_n$. By part (a), all $G_n \in \mathscr{R}$. (Note that $G_n \in \mathscr{R}_2$, which may be identified with $\mathbb{R}^2 \times \mathbb{R}$.) Part (a) applied to $B = G_n$ shows that each $x \in G_n$ is completely determined by those $y \in B$ corresponding to only countably many points $(y, y_2) \in G_n$; each such y is a coordinate, and $x \in B$ corresponds to y as above. The result follows as above.

(c) If $x \in \Omega_2$, then $\{y : (y, y_2) \in \Omega\} = \{y : (y, y_2) \in \Omega_1\}$, and the result follows from part (b).

9. Assuming the continuum hypothesis, we may identify $[0, 1]$ by the first uncountable ordinal ω_1. (This way may be $\Omega_1 = [0, \omega_1)$, $\mathscr{S}_1 = \mathscr{S}_2$ = the images of the Borel sets under the correspondence of $[0, 1]$ with ω_1, and $\mu_1 = \mu_2 = 1$ (the counting measure). Let $f = 1_D$, where $D = \{(x, y) : x = y\}$. Now $\int_0^1 f(x, y)\,dx = 1$ for each y, since $\{x : (x, y) \in D\} = \{y\}$ is countable, and for each x, $\int_0^1 f(x, y)\,dy = 0$ is considered similarly. Since f is measurable on each coordinate separately, Tonelli's theorem fails if f is measurable on each coordinate separately. Now $\int_{\Omega_2} f\,d(\mu_1 \times \mu_2) = 0$ for all x, hence $\int_0^1 \int_0^1 f\,d\mu_2\,d\mu_1 = 0$. But $\int_{\Omega_1} f\,d\mu_1 = 1$ for all y, hence $\int_0^1 \int_0^1 f\,d\mu_1\,d\mu_2 = 1$. In other words, f is not jointly measurable, for if so the iterated integrals would be equal by Tonelli's theorem.

Section 3.7

1. Let $\mathscr{S}$ be the smallest σ-field containing all measurable rectangles. Then $\mathscr{S} = \bigcup_{n=1}^{\infty} \mathscr{S}_n$, since a measurable rectangle is a measurable cylinder. But the class of sets $\mathscr{S}_n = \{\prod_{j=1}^{n} S_j : S_j \in \mathscr{S}_j\}$ is a σ-field that contains the measurable rectangles of $\prod_{j=1}^{n} \Omega_j$, and hence contains all sets in $\prod_{j=1}^{n} \mathscr{S}_j$. Thus all measurable cylinders belong to $\mathscr{S}_n$, so $\bigcup_{j=1}^{\infty} \mathscr{S}_j \subset \mathscr{S}_n$.

5.

$$\{x \in \mathbb{R}^n : f(x) = c\} = \left\{x : \sum_{i=1}^{n} x_i = 1, \quad x_i = 1, 2, \ldots, n-1, \sum_{i=1}^{n} x_i^2 = \right\}$$

$$\{x \in \mathbb{R}^n : f(x) = c\} = \left\{x : \sum_{i=1}^{n} x_i = 1, \quad x_i = 1, 2, \ldots, n\right\}$$

In each case we have a family of concentric hyperspheres or hyperellipsoidal cylinders.

Chapter 3

Section 3.2

6. (a) Let $\varepsilon > 0$, a compact subset K of Ω be given, a constant α be chosen so that the Cauchy integral formula yields

$$f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} f(\zeta) (\zeta - z)^{-1} d\zeta$$

Then for $z \in K$,

$$f(z) \frac{n!}{z^n} = \int_{\partial\Omega} f(\zeta) \zeta^{-n} d\zeta = \frac{1}{2\pi i} \int_{\partial\Omega} f(\zeta) d\zeta \int_{\partial\Omega} f(\zeta) \zeta^{-n} d\zeta$$

so

$$f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \int_{\partial\Omega} f(\zeta) \zeta^{-n} d\zeta$$

$$= \frac{1}{2\pi i} \int_{\partial\Omega} f(\zeta + z) \zeta^{-n} d\zeta$$

where Ω is the disk with center at z and radius r. An application of the Cauchy–Schwarz inequality to the functions f and ζ shows that

$$|f(z)| \leq \frac{1}{2\pi r^n} \|f\| \|\zeta^{-n}\|$$

as desired.

(b) If $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathcal{O}(\Omega)$ and $\varepsilon > 0$, then f_n converges uniformly on compact subsets to a function f and f is

on Ω. But $\mathcal{L}^2(\Omega) = L^2(\Omega, \mathcal{F}, \mu)$ where $\mathcal{F} = \{\Omega \cap B : B \text{ is a Borel set}\}$ and μ is Lebesgue measure. Hence f_n converges in L^2 to a function in $H^1(\Omega)$. Since a subsequence of $\{f_n\}$ converges to g a.e., we have $f = g$ a.e. Therefore $f \in H^1(\Omega)$ and $f_n \to f$ in L^2, that is, in the $H^1(\Omega)$ norm.

3. (a) $\|f\|_2 = 1$.

$$\frac{1}{2\pi}\int_0^{2\pi} |f(e^{i\theta})|^2\,d\theta = \frac{1}{2\pi}\int_0^{2\pi} \sum_n a_n e^{in\theta} \sum_k \bar{a}_k e^{-ik\theta}\,d\theta$$

$$= \frac{1}{2\pi}\sum_{n,k} a_n \bar{a}_k \int_0^{2\pi} e^{i(n-k)\theta}\,d\theta$$

since the Taylor series converges uniformly on compact subsets of Δ

$$= \sum_{n=0}^{\infty} a_n \bar{a}_n = 1.$$

which increases to $\sum_{n=0}^{\infty} |a_n|^2$ as r increases to 1.

(b) $\int_0^r \int_0^{2\pi} |f(z) - w|^2\,d\theta\,dr = \int_0^r r\,dr \int_0^{2\pi} |f(re^{i\theta}) - w|^2\,d\theta$

$$< 2\pi r^2 (|f| + |w|)^2 \le 2\pi \int_0^1 r\,dr = \pi M^2 (1+|w|)^2.$$

(c) $\int_0^r \int_0^{2\pi} f(z + re^{i\theta})\,d\theta\,dr = \int_0^{2\pi} \int_0^r f(z + re^{i\theta})\,dr\,d\theta = \int_0^{2\pi} \frac{2\pi}{|2\pi - z|} = 0$

so

$$\frac{1}{2\pi}\int_0^{2\pi} |f_n(re^{i\theta})|^2\,d\theta = \frac{1}{2\pi}\int_0^{2\pi} e^{2n}\,d\theta = e^{2n}.$$

hence $M(r) = 1$ for all n.

(d) If $\{f_n\}$ is a Cauchy sequence in $\mathcal{A}^1$, part (b) above shows that $\{f_n\}$ is Cauchy in $H^1(\Omega)$. By Problem 6, f_n converges uniformly on compact subsets and to $H^1(\Omega)$ to a function f analytic on Ω. Note if

$0 < \varepsilon_1 < 1$, $\varepsilon_2 > 0$; the Cauchy property of M^2 gives

$$\frac{1}{2\pi}\int_0^{2\pi} |f_n(re^{i\theta}) - f_q(re^{i\theta})|^2\,d\theta \le \varepsilon_2$$

for $n \ge n_1$ and n_1 a suitable large integer $N(\varepsilon_2)$. Letting $n \to \infty$ gives $f_n \to f$ uniformly for $|z| < r_1$

$$\frac{1}{2\pi}\int_0^{2\pi} |f_n(re^{i\theta}) - f(re^{i\theta})|^2 d\theta < \varepsilon_2 \qquad r < r_1 \qquad n > N(\varepsilon_2)$$

since r_1 may be chosen arbitrarily close to 1, $M^2(f_n - f) \le \varepsilon$ for $n \ge N(\varepsilon)$ proving completeness.

(c) Setting $z = re^{i\theta}$ corresponds to $(0, \ldots, 0, 1, 0, \ldots)$ with the 1 in position $n+1$. If the isometric isomorphism between H^2 and a subspace of l^2, the e_n are orthonormal. Now if $f \in H^2$, with Taylor coefficients a_n, $a_n = 0$, $n = 0, 1, \ldots$, then $\sum |a_n|^2 < \infty$ since f is square integrable. Thus

$$M^2(f) = \sum_{n=0}^{\infty} a_n^2 = \sum_{n=0}^{\infty} |(f, e_n)|^2$$

and the result follows from 3.2.19(f).

$$\begin{aligned}(f, z^n) &= \int_0^1 \int_0^{2\pi} f(re^{i\theta})\, r^n e^{-in\theta}\, r\,dr\,d\theta \\ &= \int_0^1 \int_0^{2\pi} \sum_{q=0}^{\infty} a_q r^q e^{iq\theta} r^n e^{-in\theta}\, r\,dr\,d\theta \\ &= \sum_{q=0}^{\infty} a_q \left(\frac{1}{2\pi}\right) \int_0^1 \int_0^{2\pi} r^{n+q+1} e^{i(q-n)\theta}\,dr\,d\theta \\ &= \begin{cases} 0, & n \ne q \\ \frac{1}{n+1}, & n = q \end{cases}\end{aligned}$$

Thus the e_n are orthonormal. Now if $f \in H^2(D)$ with

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

[illegible]

$$\int_{0}^{2\pi}\int_{0}^{2\pi} [illegible]\, dx\, dy$$

$$-\sum_{n=1}^{N} a_n \int_{0}^{2\pi}\int_{0}^{2\pi} [illegible] \left(\frac{e^{2\pi i}-e^{-2\pi i}}{2i}\right)^{n} [illegible]\, dx\, dy$$

[illegible] the Taylor series converges uniformly on compact subsets of Ω

$$-a_n \left(\frac{e^{2\pi i}-e^{-2\pi i}}{2i}\right)^{n} \int_{0}^{2\pi}\int_{0}^{2\pi} \frac{e^{i(n-k)x}}{2n+2} e^{i(n-k)y}$$

Now $f(x,y)$ is integrable on $[0,2\pi]^2$ by the Cauchy–Schwarz inequality, so we may let $N \to \infty$ and invoke the dominated convergence theorem to obtain

$$c_n = a_n \left(\frac{e^{2\pi i}-e^{-2\pi i}}{2i}\right)^{n}.$$

But the same argument with a_n replaced by $\frac{1}{2}$ shows that

$$\|f\|_{L^2} = \lim_{N\to\infty} \sum_{n=0}^{N} |a_n|^2 \int_{0}^{2\pi}\int_{0}^{2\pi} e^{i(n-k)x} e^{i(n-k)y}\, dx\, dy$$

$$= \sum_{n=0}^{\infty} |a_n|^2 \frac{2\pi}{n+1}.$$

The result now follows from [illegible]

3. (a) Let g be a continuous complex-valued function on $[0,2\pi]$ with $g(0) = g(2\pi)$. Then $g(e^{i\theta}) = h(e^{i\theta})$ where $h(z)$ is continuous on $\{z : |z| = 1\}$. By the Stone–Weierstrass theorem, h can be uniformly approximated by functions of the form $\sum_{n=-N}^{N} a_n z^n$. For the above given ε, choose N and a_n as appropriate, and define the constant function $\alpha = \frac{1}{2\pi}\int_0^{2\pi} g(\theta)\, d\theta$. Then α is the complex conjugate of each of its modulus-1 values.

This gives us uniformly approximations of those functions in C^0 by trigonometric polynomials $\sum_{n=-N}^{N} a_n e^{in\theta}$. Since any continuous function on $[0,2\pi]$ can be approximated in L^2 by a continuous function with $g(0) = g(2\pi)$, and the continuous functions are dense in L^2, the trigonometric polynomials are dense in L^2.

(b) By 3.2.5, $\int_0^{2\pi} |f(x) - \sum_{k=0}^{n} c_k e^{ikx}|^2 dx$ is minimized when $c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}dx$. Furthermore, some sequence of trigonometric polynomials converges to f in L^2 since the trigonometric polynomials are dense. The result follows.

(c) This follows from part (a) and 3.4.3 (b), or equally well from part (b) and 3.2.15 (d).

10. (a) Let $\{x_n\}$ be a fundamental sequence in M. Take $M = \{x_n : n \in \mathbb{N}\}$, where $x_n = \frac{1}{2}(1 + (-1)^n)(1 - \frac{1}{n})$. To see that M is totally bounded, for $\varepsilon > 0$ we have

$$|x_n - x_m|^2 = \left| \left(1 - \frac{1}{n}\right)x_n - \left(1 - \frac{1}{m}\right)x_m \right|^2$$
$$= \left(1 - \frac{1}{n}\right)^2 + \left(1 - \frac{1}{m}\right)^2 > 2.$$

Thus if $x_n \to x$, $x_m \to y$, then $x - y$ contains a ball of radius $\sqrt{2}$ about each $y \in M$. Since $|x_n|^2 = 1 - 1/(2n) < 1$, M has no element of maximum norm.

(b) Let M be a nonempty closed subset of the finite-dimensional space E. If $x \in N$ and $M = M \cap \{y : \|x - y\| \le \varepsilon\}$, then $N \subseteq M$ for some ε. Since $y \to \|x - y\|$, $y \in N$, is continuous and N is compact, $\|x - x_0\| = \min\{\|x - y\| : y \in N\}$ for some $x_0 \in N \subseteq M$. But the diameter of N is the same as the diameter of M; for if $x \in M$, $y \in N$ then $\|x - y_0\| \le \|x - y\|$ since $\|x - y\| \ge \|x - x_0\| - \|x_0 - y\|$. Similarly y_0 must be unique; for example, let $0 < \theta < \pi/2$, $x = (0, 1)$, $y = (\sin\theta, \cos\theta)$, $z = (-\sin\theta, \cos\theta)$.

Section 4.3

1. Since $\int_0^{2\pi} |K(x,t)|\,dt \le c$ is continuous in x, it assumes a maximum at some point $x_0 \in [0, 2\pi]$. If $K(x,t) = x(t)e^{x(t)}$, $x(t) \ge 0$, let $x(t) = e^{-x(t)}$. Let $x_1 = x$ be a sequence in $C[0, 2\pi]$ such that $\int_0^{2\pi} |x_n(t) - x_0(t)|\,dt \to 0$; we may assume that $|x_n(t)| \le 1$ for all n and t (see 4.1.7). Since K is bounded,

$$\left| \int_0^{2\pi} K(x_n, t)x_n(t)\,dt - \lim_{n\to\infty} \int_0^{2\pi} K(x_n, t)x_0(t)\,dt \right| = \lim_{n\to\infty} |(Kx_n)(x_n) - Ax_n| = 0.$$

Set $x = x_0$ in addition

$$\int_0^{2\pi} K(x_0, t)x_0(t)\,dt = Ax_0,$$

as desired.

7. (a) If $x \in \ell_1$ then

$$\|x\| = \left|\sum_{n=1}^{\infty} x_n\right| \le \sum_{n=1}^{\infty} |x_n| \le \left(\max_n |x_n|\right) \sum_{n=1}^{\infty} 1,$$

but $\sum_{n=1}^{\infty} 1 = \infty$, so we must have $\lambda = \sup\{\|x\| : \|x\|_1 \le 1\} < \infty$.

(b) Let $B = \{x : \|x\|_1 \le 1\}$. Since $\|\ \|$ is continuous, it follows that B is compact in the norm $\|\ \|_1$. Now the map $x \to \|x\|$ is a continuous real valued function on $(\ell_1, \|\ \|_1)$ and by part (a), the topology induced by $\|\ \|_1$ is weaker than the topology induced by $\|\ \|$. Thus the map is continuous on $(\ell_1, \|\ \|_1)$ and hence it assumes a minimum on B, necessarily positive since $0 \notin B$. Simply take $\alpha = \frac{1}{\lambda}$.

(c) If $x_n \to x$ in ℓ_1 and $y = \lim_{n\to\infty} x_n$ in the sup norm, then $\|x_n - y\|_1 \to 0$ by the Cauchy criterion. By (a) and Problem 6(b), $\|\ \|_1$ and $\|\ \|$ induce the same topology.

(d) By the above remarks, we must show $T = \sum_{n=1}^{\infty} x_n e_n : (x_1, x_2, \ldots) \to x$ is a one-to-one, onto, linear, bicontinuous map of ℓ_1 and ℓ^∞. Indeed, $\|\sum_{n=1}^{\infty} x_n e_n\|$ is the Euclidean norm of $(x_1, \ldots, x_n, 0, \ldots)$ in ℓ^∞. If $x_n \to x$ in ℓ^∞, then $x_n - x_m \to 0$ as $n, m \to \infty$ because $\|x_n - x_m\| \le \|x_n - x\| + \|x - x_m\|$. Thus $\{T x_n\}$ is a Cauchy sequence in ℓ^∞. If $T x_n \to y \in \ell^\infty$, then $x_n = T^{-1} T x_n \to T^{-1} y$.

8. For $x \in \ell^\infty$ we use $\{\ell_1$ and Problem 7; if $\{x\}$ holds, then $\{x : \|x\| \le r\}$ is compact for small enough $r > 0$. Because every closed set is compact, we see that the map $x \to \|x\|$ is a homeomorphism. But since closed bounded sets in ℓ^∞ are not compact, $\{x\}$ must have to be compact.

To prove that (f): suppose $\{x_n\}$ is a subsequence of $\{x_n\}$ such that $\|x_n\| = 1$. Suppose we have chosen $x_1, \ldots, x_{n-1} = 1$ and such that $\|x_i - x_j\| \ge \frac{1}{2}$ for $i = 1, \ldots, n-1$; $j \ne i$. If L is not finite dimensional, then $\{x_1, \ldots, x_{n-1}\}$ is a proper subspace of L, necessarily closed by Problem 7(d). By Problem 8, we can find $x_n \in L$ with $\|x_n\| = 1$ and $\|x_i - x_n\| \ge \frac{1}{2}$, $i = 1, \ldots, n-1$. This construction $x_1, x_2, \ldots$ satisfies $\|x_n\| = 1$ and $\|x_n - x_m\| \ge \frac{1}{2}$ for $n \ne m$; hence the unit sphere cannot possibly be compact by Problem 8. Thus L must be finite dimensional.

11. (a) Define $\lambda(A) = \mu(\{x : x \in [0,1],\ (x, x_1, x_2, \ldots) \in A\})$. If $A_1, A_2, \ldots$ are disjoint sets in $\mathcal{A}$ whose union is A, then $\lambda(A) = \sum_{n=1}^{\infty} \lambda(A_n)$ since μ is countably additive and $\sum_{n=1}^{N} \lambda(A_n) \longrightarrow \lambda(A)$. Moreover

$$\int_0^1 \sum_{n=1}^{\infty} f_n(x, x_1, x_2, \ldots)\, dx = \sum_{n=1}^{\infty} \mu(A_n) = \lambda(A)$$

for finite sums of χ_E. Thus $\int_0^1 |f_n(x)| dx = 0$ and $f_n \to 0$ a.e. $[\mu]$, so we may write $f_n^{2^n} \to 0$ and use the continuity of T to obtain $Tx = 0$. By the Banach–Alaoglu theorem, we have $\|x\| \geq \|y\|$ where y is some weak-star limit of x. Thus $\int f dx = \int_0^1 f(y) dy$ when f is continuous; hence when f is a simple function. Since f is continuous, y is μ integrable, and the linear valued set-up functions are dense in L^1, the result holds when x is a bounded Borel measurable function.

Since $\{y_1, y_2, \ldots\}$ is a countable set of finite-valued simple functions increasing to y, we have

$$\int_0^1 y^2 dx \geq \int_0^1 y_n^2 dx = \int_0^1 y_n^{2^n} dx = \int_0^1 y^{2^n} dx$$

$$= \int_0^1 y^{2^n} dx \quad \text{since} \quad y^{2^n} \text{ is bounded}$$

$$\leq \|f\| \int_0^1 |f(x)| dx = \int_0^1 |f(x)| dx \quad \text{since} \quad \|f\| = 1 = \|f\|_1$$

Thus $\|y\|_2 \leq \|f\|_1$ hence by the monotone convergence theorem, $\|y\|_2 \leq \|f\|_1$ in particular $y \in L^2$. But now Hölder's inequality and the fact that finite-valued simple functions are dense in L^1 yield $\int f(x) y dx = \int_0^1 y dx$ for all $f \in L^1$. Hölder's inequality also gives $\|f\|_1 \leq \|y\|_2$; hence $\|f\| = \|y\|_2$. If y_1 is another such function, then $\int f(y - y_1) dx = 0$ for all $f \in L^1$. By the above equation, $\|y - y_1\|_2 = 0$, so $y = y_1$ a.e. $[\mu]$.

(b) (i) If $\sup_n \|x_n\| = \infty$, we can find a subsequence $\{x_{n_k}\}$ with $\|x_{n_k}\| \geq k^2$; let $x = \sum k^{-1} x_{n_k}$; then $\|x\| \leq \sum k^{-2} < \infty$, hence $\int_0^1 |x(t)| dt < \infty$, yet

$$\int_0^1 |x_{n_k}| dt \geq k^2 (k^{-1}) = k$$

But then $\|x\| = \infty$.

(ii) Since $\sup_n \|x_n\| = \infty$, as in (i), we have

$$\|x_n\|_2^2 \geq \|x_n\|_2 \|x_n\|_2 = \|x_n\|_2 \geq 1 \int_{x_n \geq 1} |x_n|^2 dx$$

Since $\|x_n\|_2 \to \infty$ as $n \to \infty$, so does $\|x_n\|_2 \|x_n\|_2$; it follows that $\int_{x_n \geq 1} |x_n|^2 dx \to \infty$ as $n \to \infty$. By symmetry we may assume that $x_n \geq 0$. Then

$$\int_{x_n \geq 1} |x_n - x_n|^2 dx = \int_{x_n \geq 1} x_n^2 dx$$

$$= \int_{x_n \geq 1} x_n^2 dx \to \infty$$

as $t \to \infty$. Thus g_n converges to 0 for each x and since $|g_n(x)| \le |f(x)|$ for all x, $\int |g_n| \to 0$. If $\{g_{n_k}\}$ is another sequence of this form with $\|g_{n_k}\|_1 \to L$, the above argument with h_n replaced by g_{n_k} shows that $\|g_{n_k} - g_n\|_1 \to 0$, hence $g_{n_k} \to g$ a.e.

(iii) Let $A \subset \mathbb{R}$, $\chi(x) = \cos 2\pi x$. In (ii) we may take all $A_n \supset A$ so that $\alpha_n = \alpha$ and $\omega_n = \omega$. Since $\alpha = \omega$ and $\cos = \sin$, $\hat{f}(\omega) = \int_A \cos(\omega x - \alpha)\,dx = \int_A \chi(\omega x)\,dx$. It follows that $\hat{f}(\omega) = \hat{f}(\omega + 1)$ for all ω in the sequence. If $\hat{f}(\omega) = \infty$, then ω must be in $\mathbb{R}$ since $\omega \in \mathbb{Z}$. Since $\omega \notin \mathbb{Z}$, the orthogonality of χ and Hölder's inequality extend this result to all $\omega \in \mathbb{R}$.

(c) The argument of (b) yields a χ_n integrable f such that $\hat{f}(\omega) = \hat{\chi}_n(\omega)$ for all bounded Borel measurable ω. Let $B = \{\omega : |\omega| \le N\}$. Then

$$\begin{aligned} \int_B |\hat{f}|^2 &= \int_B |\hat{f}|\,|\hat{f}| = \int_B \left(\int e^{-2\pi i \omega x} f(x)\,dx\right) \overline{\hat{f}(\omega)}\,d\omega \\ &= \int f(x) \left(\int_B e^{2\pi i \omega x} \overline{\hat{f}(\omega)}\,d\omega\right) dx \le \|f\|_1 \sup_{x \in \mathbb{R}} \left|\int_B e^{2\pi i \omega x} \overline{\hat{f}(\omega)}\,d\omega\right|. \end{aligned}$$

Thus if $C > \|f\|_1$ we have $\int_B |\hat{f}|^2 \le C$ whenever $\int |\chi_n|^2 \le C^2$. As in (b), we obtain $\hat{f}(\omega) = \hat{\chi}_n(\omega)$ for all $\omega \in [-N, N]$ with $N = \|\chi_n\|_2^2$, and $\hat{f}$ is locally χ_n a.e.

(d) Part (c) of (b) holds with the same proof. Since $\hat{\chi}$ is the sum of distinct sets A_n with $\chi(\omega_n) = \alpha_n$, we can define f as in (b) by taking $\omega = \omega_n$ on A_n. Since $|\alpha_n| \le 1$ for all n, we now have $\hat{f} \in L^\infty$ and $\|\hat{f}\|_\infty \le \|f\|_1$. If $\omega \in A_n$, then $\sum_{k=1}^{n-1} \alpha_k \frac{e^{i\omega}}{\omega - \omega_k}$ is finite and

$$\begin{aligned} f(x) &= \sum_{n=1}^{\infty} \frac{1}{n^2} \hat{\chi}_n(x) = \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{A_n} \chi(\omega x)\,d\omega \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2} \int_{A_n} e^{i\omega x}\,d\omega = \int_{\mathbb{R}} e^{i\omega x}\,d\omega. \end{aligned}$$

Since $|e^{ix}| \le \|f\|_\infty$ by Hölder's inequality (with $p = 1$, $q = \infty$), the result follows.

Section 5.4

1. Let $\{x_n\}$ be a Cauchy sequence in M, and set ε be any element of L with norm 1. By 5.4.5 (c), there is an $N \in \mathbb{N}$ with $\|x\| = \|x\|\,\varepsilon$ and $f(\varepsilon) = \|x\|$. Now define $A_n = [0, N]$ by $A_n x = f(x)\varepsilon$. Then $\|x_n - A_n x\| = |f(x)|\,\|x_n - x\| = \|x - x_n\| \to 0$ and $\|A_n - A_m\| \le \|x_n - x_m\|$.

$\varepsilon > 0$. By hypothesis, the A_n converge uniformly to zero on $L = [0, M]$, thus $|A_n(x) - A_n(y)| < \varepsilon$ for $|x - y| < \delta$ and $n \geq N$. Then $|x_n - x_{n+1}| = |A_n x_{n-1} - A_{n+1} x_n| \leq |A_n - A_{n+1}| < \varepsilon$.

5. Let V be the set of all bounded scalar-valued functions on $[0, \infty)$ with the sup norm, and let M be the subspace of V consisting of step functions

$$s = \sum_{i=1}^{n} c_i \chi_{[a_i, b_i)},$$

where the $[a_i, b_i)$ are disjoint and $\subset [0, T]$. Define μ on M by

$$\mu(s) = \sum_{i=1}^{n} c_i (b_i - a_i).$$

Now $|\mu(s)| \leq \max_i |c_i| \sum_{i=1}^{n} (b_i - a_i) \leq \max_i |c_i| \mu([0, T]) = \mu(M) \|s\|$; hence $|\mu| \leq \mu([0, T]) < \infty$. By the Hahn–Banach theorem, μ has an extension to a continuous linear functional λ on V with $\|\lambda\| = \|\mu\|$. Define $\alpha(t) = \lambda(\chi_{[0,t)})$, $t \geq 0$. Since λ is linear, α is finitely additive, and since λ is an extension of μ, α is an extension of m. Now $|\alpha(t)| \leq \|\lambda\| \cdot \|\chi_{[0,t)}\| = \|\lambda\| \cdot \mu([0, t)) \leq \|\lambda\| \mu([0, T])$.

But $\|\chi_{[0,t)}\| = 1$, so that $|\alpha| \leq \mu([0, T]) < \infty$, a contradiction.

6. Since $A_n x \to Ax$, we certainly have $\sup_n \|A_n x\| < \infty$. By the uniform bounded nessprinciple $\sup_n \|A_n\| = M < \infty$, hence

$$\|Ax\| = \lim_{n \to \infty} \|A_n x\| = \lim_{n \to \infty} \|A_n x\|$$
$$\leq \|x\| \liminf_{n \to \infty} \|A_n\| \leq M \|x\|.$$

12. Let L be the set of all complex-valued functions on $[0, 1]$ with a continuous derivative on $[0, 1]$, let M be the subset of all x in L such that $x(0) = 0$, let V be the space of all complex-valued functions on $[0, 1]$ with the sup norm, and let A be the mapping $x \to x'$, $x \in L$. Then A is a linear map of L onto M, and A is closed. That $x_n \to x$ and $x_n' \to y$ then simply says that the sequence $\{x_n\}$ converges to x and the sequence $\{x_n'\}$ converges uniformly to y. Thus $x(t) = \lim_{n \to \infty} x_n(t) = \lim_{n \to \infty} x_n(0) + \int_0^t x_n'(s)\,ds = x(0) + \int_0^t y(s)\,ds$. Thus $x' = y$, so that A is indeed closed. Now $\|x\| = \sup_{0 \leq t \leq 1} |x(t)| = \sup_{0 \leq t \leq 1} |\int_0^t x'(s)\,ds| \leq \|x'\|$.

13. By the open mapping theorem, A is open; hence $A(\{x \in L : \|x\| < 1\})$ is a neighborhood of 0 in M; say $\{y \in M : \|y\| < \delta\} \subset A(\{x \in L : \|x\| < 1\})$. If $y = Ax$, $y \neq 0$, then $\delta y / 2\|y\|$ has norm smaller than δ, hence equals Ax for some $x \in L$, $\|x\| < 1$. Let $x = (2\|y\|/\delta)x$; then $Ax = y$ and $\|x\| < \delta \|x\|/\delta$. Thus we may take $k = 2/\delta$.

CHAPTER 4

Section 4.8

7. To prove (a), let $A, B \in \mathcal{A}$. If $A \subset B$, then $B = A \cup (B \setminus A)$, so the assumed finite additivity implies

$$P(B) = P(A) + P(B \setminus A) \geq P(A),$$

$$P(B) = P(A) + P(B \setminus A) \leq P(A) + \sum_{k=1}^{n} P(C_k),$$

$$P(A) = P(A \cap C_1) + \cdots + P(A \cap C_n).$$

Thus $A \cap B$ can be added to the finite class $\mathcal{C}$ without destroying independence. Since $A_1 \cap A_2 \cap \cdots \cap A_n$ belongs to $\mathcal{C}$, (b) follows. The proofs of (b) and (c) are quite similar, and (d) follows from (a), (b), and (c).

Now let $\mathcal{C}_1 = \{A : A \cap B \in \mathcal{C}\}$, where A and C are independent and B and C are arbitrary sets. Since $P(A \cap B \cap C)$ and $P(A \cap B)P(C)$ [see 4.3.5(b)], $A \cap B$ could be added to $\mathcal{C}_1$.

Finally, we show that the $\sigma(\mathcal{C}_i)$ are independent iff each $\mathcal{C}_i$ is closed under finite intersections. First, suppose the collection of classes $\mathcal{C}_i$ such that $\mathcal{C}_i$ is closed under finite intersections $\mathcal{C}_1, \ldots, \mathcal{C}_n$ and $A_i \in \sigma(\mathcal{C}_i)$ and $A_j \in \mathcal{C}_j$, $j \neq i$, are independent. Partially order the $\mathcal{C}_i$ by inclusion. Each chain has an upper bound, by the union of the classes in the chain, so there is a maximal class $\mathcal{C}_i'$. Since $\mathcal{C}_i$ is closed under finite intersections, it is closed under arbitrary differences by [see 4.3.5(b)], hence by (a) and (b), $\mathcal{C}_i'$ is a π-field. Thus $\sigma(\mathcal{C}_i') = \mathcal{C}_i'$, and consequently $\mathcal{C}_i$ can be replaced by $\sigma(\mathcal{C}_i)$ while preserving independence. Since i was arbitrary, the result follows.

Section 4.9

4. $P(X = k) = P(X = k^{-1}(S)) = \int_{k^{-1}(S)} f(x)\,dx$. Under the transformation $x = h(y)$, $dx = g(y)\,dy$, this becomes $\int_S f(h(y))|g(y)|\,dy$ [see Problem 6, Section 2.3]. The result follows.

6. Let $A = \{X_1 = 2 + Y_1\}$, $Y_1 = \min(X_2, X_3)$. Then $X_1 = \sum_{j=1}^{3} jI_{A_j}$, where I_A is the indicator of A. Therefore $E(X_1) = \sum_{j=1}^{3} jE(I_{A_j}) = \sum_{j=1}^{3} jP(A_j)$, where $P(A_1) = \int_A \tau_{1,2,3}$. By symmetry,

$$E(X_1) = nP(X_1 = 2 + Y_1)$$

$$= n\sum_{k=1}^{n} P(Y_1 = k, X_1 = 2 + Y_1)$$

$$
\begin{aligned}
&= P(X_1 \le x_1) P(X_2 \le x_2 \mid X_1 \le x_1) \\
&\quad \times P(X_3 \le x_3 \mid X_1 \le x_1, X_2 \le x_2) \cdots P(X_n \le x_n \mid X_1 \le x_1, \ldots, X_{n-1} \le x_{n-1}) \\
&= F_1(x_1) \int_{-\infty}^{x_2} f(x_1, x_2)\, dx_2 \Big/ \int_{-\infty}^{x_1} f(x_1, x_2)\, dx_2 \int_{-\infty}^{x_3} f(x_1, x_2, x_3)\, dx_3 \Big/ \int_{-\infty}^{x_2} f(x_1, x_2)\, dx_2 \\
&\quad \times \cdots \int_{-\infty}^{x_n} f(x_1, \ldots, x_n)\, dx_n \Big/ \int_{-\infty}^{x_{n-1}} f(x_1, \ldots, x_{n-1})\, dx_{n-1} \\
&= F_1(x_1) \int_{-\infty}^{x_2} f(x_1, x_2)\, dx_2 \Big/ f_1(x_1) \int_{-\infty}^{x_3} f(x_1, x_2, x_3)\, dx_3 \Big/ \int_{-\infty}^{x_2} f(x_1, x_2)\, dx_2 \\
&\quad \times \cdots \int_{-\infty}^{x_n} f(x_1, \ldots, x_n)\, dx_n \Big/ \int_{-\infty}^{x_{n-1}} f(x_1, \ldots, x_{n-1})\, dx_{n-1}.
\end{aligned}
$$

7 Define F^{-1} and x_0 as suggested. Then if $0 < y < 1$, $F^{-1}(y) \le x$ iff $y \le F(x)$. For if $F^{-1}(y) = x_1$ we can find $y_n \downarrow y$ such that $F(x_n) < y_n$ and thus $F(x) \ge F(x_n) > y$. If $F^{-1}(y) = x_0$ it is right-continuous, so we can find $x_n \downarrow x$ such that $F(x_n) < y$, therefore $F^{-1}(y) \ge x_n > x$. Now

$$
P[\text{for } X(\omega) \le x_0 = P[\text{for } F^{-1}(\omega) \le x] = F(x), \quad \text{for } y \in [0,1] \quad P[Y \le y] = F(y).
$$

Thus also shows that X is measurable, and that

$$
X(\omega) = \min\{x : F(x) \ge \omega\}, \qquad 0 < \omega < 1.
$$

Section 4.IV

2 Suppose F is $\equiv$ Borel and absolutely continuous, $F = F_1 + F_2$, where

$$
F_1(x) = \begin{cases} 0, & x < 2, \\ \frac{1}{2}, & x \ge 2, \end{cases}
$$

$$
F_2(x) = \int_{-\infty}^{x} f_2(t)\, dt
$$

where

$$
f_2(x) = \begin{cases} 0, & x < 2 \text{ or } x > 3, \\ \frac{1}{2}, & 2 < x < 3. \end{cases}
$$

Then

$$
\begin{aligned}
\sigma(F) &= \int_{-\infty}^{\infty} x\, dF(x) = \int_{2}^{\infty} x^2\, dF_1(x) = \int_{-\infty}^{\infty} x^2\, dF_2(x) \\
&= 2^2 \left(\frac{1}{2}\right) + \int_{-\infty}^{\infty} x^2 f_2(x)\, dx \\
&= \frac{4}{2} + \frac{1}{2} \int_{2}^{3} x^2\, dx = \frac{20}{6}.
\end{aligned}
$$

Section 4.4

1. (a) Assume $X_n \in \mathcal{L}(\mathcal{F}^X_n)$ and $\mathcal{F}^X_n \subset \mathcal{F}^X_{n+1}$, $n = 1, 2, \ldots$. By properties

$$P[X_{n+1} \in A_{n+1}, \ldots, X_{n+k} \in A_{n+k} \mid \mathcal{F}^X_{n+1}] \in \mathcal{L}$$

$$= P[X_{n+1} \in A_{n+1} \mid \mathcal{F}^X_{n+1}] \cdots P[X_{n+k} \in A_{n+k} \mid \mathcal{F}^X_{n+1}]$$

where $A_{n+1}, \ldots, A_{n+k}$ are measurable rectangles in $\prod_{i=1}^{n+1} S_i$, and $\mathcal{L}$ is the monotone class of measurable rectangles in $\prod_{i=1}^{\infty} S_i$. (For the formula is true when $A_{n+1}, \ldots, A_{n+k}$ are finite disjoint unions of measurable rectangles. Then approximate $\mathcal{L}$ by the monotone class theorem.) It follows that for all $A \in \prod_{i=1}^{\infty} \mathcal{S}_i$ and $B = \prod_{i=1}^{n} S_i$

(b) Let $\mathcal{L}$ be the class of sets $A \in \mathcal{F}^X$ such that

$$P[\{X_1, X_2, \ldots\} \in A \mid B] = P[\{X_n, X_{n+1}, \ldots\} \in A \mid B]$$

Since the X_i are independent and $\mathcal{L}$ contains all finite dimensional measurable rectangles, hence $\mathcal{L}$ contains all finite dimensional measurable rectangles. But $\mathcal{L}$ is a monotone class hence $\mathcal{L} = \mathcal{F}^X$ and the result follows.

Chapter 5

Section 5.1

1. (a) The conditional density of X given Y is $f(x|y) = f(x,y)/f_Y(y)$, where $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$ is the marginal. Thus

$$E[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f(x,y)/f_Y(y)\,dx$$

assuming $E[|g(X)|]$ exists [cf. 5.3.5(a)]. By (5).

(b) $E[I_A] = P[Y \in A] = \int_A f_Y(y)\,dy$ [see 5.3.5(a)]. Also

$$\int_B g(y) f_Y(y)\,dy = \int_{-\infty}^{\infty} \int_B g(y) f(x,y)\,dy\,dx$$

and

$$E[X I_A(Y)] = \int_{-\infty}^{\infty} \int_A x f(x,y)\,dy\,dx$$

$$= \int_A \int_{-\infty}^{\infty} x f(x,y)\,dx\,dy.$$

(c) $E\,X(t) = E(U_1)\sqrt{\pi t} = 0$, because

$$F(y) = \iint_{x^2+y^2\le 1} f(x,y)\,dx\,dy$$

and

$$E(XY) = \iint_{x^2+y^2\le 1} xy\,f(x,y)\,dx\,dy.$$

2. By 5.2.1, $P[X \in A, Y \in B] = \sum_{k\in B} P[X \in A]\int_k f(y)\,dy$. Then take $A = \mathbb{R}$ to see that Y has density $f(y) = \sum_{k\in I} P[X = k] f(y)$. Since in general

$$P[X \in A] = \sum_{k\in I} P[X = k]\,P[Y = k],$$

where $P[X = k]$ is as specified in the problem, then

$$\begin{aligned}\int_A P[X \in A]\,f(y)\,dy &= \int_A P[X \in A]\,f(y)\,dy \\ &= \int_A \sum_{k\in I} P[X = k]\,f(y)\,dy \\ &= \sum_{k\in I} P[X = k]\int_A f(y)\,dy \\ &= P[Y \in A,\ Y \in B].\end{aligned}$$

The result follows from 5.3.1.

Section 5.4

1. (a) We show that $\{X(\omega) : X(\omega) = 0\}$ is a function. If any then be removed arbitrarily off $X(\omega)$. If we do not have a function, then there are points $\omega_1, \omega_2 \in \Omega$ with $X(\omega_1) = X(\omega_2)$ but $Z(\omega_1) \ne Z(\omega_2)$. Now $Z(\omega_1) = Z^{-1}(X(\omega_1)) = B_1$ with $X(\omega_1) \in C_1$, $X(\omega_2) \in C_2$. Since $Z^{-1}(X(\omega_2)) = Z^{-1}(X(\omega_1)) = B_1$, and since $C_1, C_2 \in \mathcal{S}^1$, we have $Z^{-1}(C_2) = Z^{-1}(B_1)$ for some $B_1 = \mathcal{S}^1$, $j = 1, 2$. Now, since $X(\omega_2) \in C_2$, there is $\omega \in X^{-1}(C_2) = X^{-1}(B_2)$. Therefore $X(\omega) \in B_2$. But $X(\omega) \in B_1$ by definition, $X(\omega) = x_2$ as well as x_1. Since only $X(\omega) = B_1, B_2$ $\in \mathcal{S}^1$. But $X(\omega_1) = X(\omega_2)$, and this contradicts the assumption $B_1 = B_2$ $\in B_1 \cap B_2$ is always empty.

(b) Let $\Omega_0 = X(\Omega)$. If $C \in \mathcal{S}^1$, then

$$\text{[illegible]}$$

But

$$\mathcal{Z} \quad (C_j = Y^{-1}(B_j) \quad \mathcal{T}^{-1}(S \cap \Omega_1) \qquad \text{for some} \qquad B_j \in \mathcal{F}'$$

Since X maps onto Ω_2, $X[X^{-1}(A)] \quad A$ for any $A \subset \Omega_2$; hence $f^{-1}(C)$

$\mathcal{T}\Omega_2 = B \cap \Omega_2 \in \mathcal{F}'$ But if $f(S)^{-1} \quad \Omega_1 = \omega$, then

$$f^{-1}(C) \cap \Omega_1 = \begin{cases} \omega & \text{if } \omega \in C \\ \Omega - \Omega_1 & \text{if } \omega \notin C \end{cases}$$

Therefore $f^{-1}(C) \cap \Omega_1 \in \mathcal{F}$; hence $f^{-1}(\mathcal{F}') \subset \mathcal{F}$

Section 5.5

$$\int_{\text{Set}\,\Omega\times\Omega} Y\,dP = E[Y\chi_{(\omega,\infty)}(X)]$$

$$= E[Y]E[\chi_{(\omega,\infty)}(X)] \qquad \text{by independence}$$

$$= \int_{\Omega_2} y\,dF_Y(y) \int_{(\omega,\infty)} dF_X(x)$$

$$= \int_{\Omega_2} \int_{(\omega,\infty)} y\,dF_X(x)\,dF_Y(y)$$

$$= \int_{\text{Set}\,\Omega\times\Omega} Y\,dP \qquad \text{by independence}$$

$$= \int_{(\omega,\infty)\times\Omega_2} y\,dF(x,y)$$

Thus

$$\int_{A\times\Omega_2} y\,dF = \int_{\omega\in A} E[Y|X=x]\,dF_X \qquad (1)$$

for C a measurable rectangle $A \times B$, $A \subset \mathcal{F}'$, $B \subset \mathcal{F}'$ (where $X : (\Omega, \mathcal{F}) \to (\Omega_2, \mathcal{F}')$, $Z : (\Omega, \mathcal{F}) \to (\Omega_1, \mathcal{F}')$). By the monotone class theorem (1) holds for all $C \in \mathcal{F}' \times \mathcal{F}'$. [Impossible if Y is not in above set of C's which for $C = \mathcal{F}' \times \Omega_2$ and $C_2 = \Omega_1$, then it holds for $\mathcal{F}'$]

Section 5.6

1. (a) Let $\varphi(x,y) = \chi_A(x)\chi_B(y)$ be the indicator of a measurable rectangle $A_1 \times \Omega_2 \in \mathcal{F}' \times \mathcal{F}'$. Then

$$\int_{A\times B} \varphi(X,Y)\,dP = P[X \in A \cap A_1, Y \in B]$$

$$= \int_{A\cap A_1} P[Y \in B, X = x]\,dP_X(x)$$

$$= \int_A \chi_{A_1}(x)\,\chi_B(y)\,dP(x,y)$$

Thus

$$E[g(X,Y)|X=x] = \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y|x)\,dy$$
$$= \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dy \qquad \text{for } x \in S_X.$$

Thus the result holds for g of this type. By the usual method of indicators of arbitrary sets in $\mathcal{B}^1 \times \mathcal{B}^1$ using the monotone class theorem, we then extend to arbitrary g in the usual way.

11: We have

$$F_{X,Z}(x,z) = P(X \le x, Y - X \le z) = E[I_{(-\infty,x]}(X) I_{(-\infty,z]}(Y-X)]$$
$$= \int_{-\infty}^{\infty} F_{Y|X}(z+x,y)\,dF_X(x)$$
$$= F_X(x) F_Y(z) \qquad \text{since } X \text{ and } Y \text{ are ind.}$$

12: $P(X=x|Y=y) = P(X=x, Y=y)/P(Y=y)$ by elementary conditional probability. The result follows from (5).

13: (a) Since $\mu(E) < \infty$ and $\lambda < 1$, there is an open set $V \supset E$ such that $\mu(V) < \lambda^{-1}\mu(E)$. If V is a disjoint union of open intervals I_n, then

$$\lambda \sum_n m(I_n) = \lambda m(V) \ge \mu(E) = \mu(E \cap V) = \sum_n m(E \cap I_n). \quad (*)$$

Therefore $\lambda m(I_n) \ge m(E \cap I_n)$ for some n. Note that

$$\sum_n m(I_n) = \mu(V) < \lambda^{-1}\mu(E) < \infty$$

so it is not possible to have both signs included in $(*)$.

(b) By (a) there is an open interval I such that $\mu(E \cap I) \ge \frac{1}{2}\mu(I)$. We show that $I = (\mu(I), \mu(I) + m(I))$. Let $x \in (\mu(I), \mu(I) + m(I))$. If $x \notin \mu(E \cap I)$, then x is a limit point of the measure of $E \cap I$, so there is $\delta > 0$ such that $\mu(E \cap I) \le \frac{1}{2}m(I)$. But $\mu(E \cap [\mu(I), x]) \ge \mu(E \cap I) + \delta$, an open interval of length less than $m(I)$. $\frac{1}{2}m(I) = \frac{1}{2}\mu(I)$, a contradiction. Thus $x \in \mu(E \cap I)$ for some $x \in E$. Hence $x = \mu(z)$, an identical

(c) Since the circle is compact, there is a subsequence converging to a point on the circle. Given any positive integer N, choose n_N such that $n \geq N$ and $x_n = (\cos \alpha, \sin \alpha)$ implies that $x_{n+1} = (\cos(\alpha + 2\pi/N), \sin(\alpha + 2\pi/N))$. Then $0 < x_{n+1} - x_n < 2\pi/N$. Note that $x_{n+1} \neq x_n$ since $\pi/2\pi$ is irrational. Thus $x_{n_N}, x_{n_N+1}, \ldots, x_{n_N+N}$ is a chain that connects x_{n_N} to its antipode on the circle, with the distance between successive points less than $2\pi/N$. Thus, given $\varepsilon > 0$, we can find $n_N \geq N$ such that $|x_{n_N} - x_n| < \varepsilon$. The result follows.

(d) Since $C = \{(x, y) \mid x = 0 \text{ or } y \in S\}$, it suffices to consider E. But S is dense in the set of nonnegative real numbers, so S contains a sequence $\{s_n\}$ converging to 0. Equivalently, consider $\{(s_n, 0) \mid n \geq 1\}$ in an open ε-disk in the circle. Note is irrational. But in this case C is also connected, and the result follows from (c).

(e) Let $A \subset \mathcal{B}(\mathbb{R}^n)$ be $\mathcal{M}$. We assume $m(A) > 0$. Let $\varepsilon > 0$. Since A is measurable, there exists a finite collection $\{E_1, \ldots, E_k\}$ of disjoint open cubes such that $m(A \Delta \bigcup_{i=1}^k E_i) < \varepsilon$. Let $E = \bigcup_{i=1}^k E_i$. Then $m(A \cap E) > 0$. Since E is open, there exists a cube Q such that $m(A \cap Q) > \frac{1}{2} m(Q)$. Let x be the center of Q. Then $Q = x + [-r, r]^n$ for some $r > 0$. Let $B = A \cap Q$. Then $m(B) > \frac{1}{2} m(Q)$. Let $y \in B$. Then $y = x + z$ for some $z \in [-r, r]^n$. Thus $B \subset (B - z) \cup (B + z)$. Since $B \subset Q$, $B - z \subset [-2r, 2r]^n$. Thus $m(B - z) \leq 2^n m(Q)$. Hence $m(B \cap (B - z)) \geq m(B) - m(B - z) > \frac{1}{2} m(Q) - 2^n m(Q) = (\frac{1}{2} - 2^n) m(Q)$. Since $n \geq 1$, $\frac{1}{2} - 2^n > 0$. Thus $B \cap (B - z)$ is nonempty. Let $w \in B \cap (B - z)$. Then $w \in B$ and $w + z \in B$. Hence $w, w + z \in A$. Thus $z \in A - A$. But $z \in [-r, r]^n$. Hence $A - A$ contains a neighborhood of the origin. Now assume $m(E) < \infty$. E is the union of a sequence of disjoint cubes $\{Q_i\}$ so that $\sum m(Q_i) < \infty$. Since $\sum m(Q_i)$ converges, $m(Q_i) \to 0$. Each cube Q_i contains a ball B_i of radius r_i such that $m(B_i) > \frac{1}{2} m(Q_i)$. Thus $m(E) < \infty$. It follows that if E_1 is Lebesgue measurable, then $m(E_1) < \infty$.

Now if $x \in \mathbb{R}$, then x is equivalent to some $y \in [0, 1)$, since $x - y \in \mathbb{Z}$. Therefore $\mathbb{R} = \bigcup_{r \in \mathbb{Q}} (E_1 + r)$, $r \in \mathbb{Q}$. But if $x = y + r$, where $y, z \in E_1$, $x - y = r \in \mathbb{Q}$. Note that $r \neq 0$ (up to addition). Thus $y - z = r$, but since $y, z \in E_1$, $y = z$ and therefore $r = 0$. Thus the sets $E_1 + r$, $r \in \mathbb{Q}$, are disjoint.

Now, E_1 is measurable. Then $\mu(E_1) = m(E_1)$ by translation invariance of Lebesgue measure. Since μ is countably additive, the preceding paragraph implies that $\mu(\mathbb{R}) = 0$ or ∞, a contradiction.

(f) Let $x \in \mathbb{R}$. Then $x = y + r$ for some $y = F(x) \in A$ and $r \in \mathbb{Q}$. Since $A = B \cup C$, it follows that $x = y + r \in B \cup C$. Let B' be the subset of B consisting of those x such that $x = y + r$ where $y \in C$ and $r \in \mathbb{Q}$. Then $x = y_1 + r_1 = y_2 + r_2$, where $y_1, y_2 \in C$ and $r_1, r_2 \in \mathbb{Q}$. It follows that $y_1 - y_2 = r_2 - r_1 = q \in \mathbb{Q}$. Hence $y_1 = y_2 + q$. But then $x = y_1 + r_1 = y_2 + q + r_1 = y_2 + r$, where $r = q + r_1 \in \mathbb{Q}$. Since x is arbitrary, this same argument applies to C. It follows that $B' = C$. Hence, since $B' = A - \mathbb{Q} \subset M$, any Borel subset of M has Lebesgue measure 0, as translates are invariant.

(g) The first statement follows from (f) since $F^{-1}(A) \subset C \subset F^{-1}(A) \cup N$. So the second statement follows from (f) since

4. (a) If $\bigcup_{i=1}^n (B_i \cap A) = \bigcup_{i=1}^n (B_i \cap A')$, then $B_1 \cap A = B_1 \cap A'$. If $B_1 \cap A = B_1 \cap A'$, say, $\mu(B_1 \cap A) = \mu(B_1 \cap A') > 0$, then $B_1 \cap A'$ is not a subset of A' since A' has zero Lebesgue measure. If

$\mathcal{M}$ then is not a π-system, contradicting $\mathcal{B} = \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$. Thus $\mu(B) = \mathcal{B}(\mathbb{R}) = 0$; a symmetrical argument shows that $\mu(\mathcal{B}) = \mu(\mathbb{R}) = 0$.

(c) If $B \in \mathcal{C}$,

$$\mu(B) = \frac{1}{2}\mu(B) = \frac{1}{2}\mu(B) = \int_B \frac{1}{2}\,dx$$

Thus $\mu(B) = \frac{1}{2}\lambda(B)$. But $\mu(B) = \mu(B \cap [0,1]^2)$, hence $\mu(B) = \frac{1}{2}\lambda(B \cap [0,1]^2)$.

(d) If $B_1, B_2 \in \mathcal{B}(\mathbb{R}^2)$ with $B_1 \subset B_2$ and $\lambda(B_2) = 0$, then $\lambda(B_1) = 0$ as well.

(e) By exercise (d), there is a set $B' \in \mathcal{B}$ with $P(B') = 0$ such that for every $x \in B$, $Q(x, B') = Q(x, \mathbb{R}) = 1$ and $Q(x, B') = \lambda(x)$ for $x \in B'$. Therefore B with rational endpoints. For any such $Q(x, \cdot)$, $Q(x, \cdot) = 1$. Consider a sequence of rational intervals decreasing to $\{x\}$ and note that $Q(x, \cdot)$ is a probability measure. For $x \in B$, then $Q(x, \{x\}) \geq Q(x, B) = \frac{1}{2}$, and if $x \in B$, then $Q(x, \{x\}) \geq Q(x, B') = \frac{1}{2}$, a contradiction.

CHAPTER 6

Section 6.1

1. (a)
$$\frac{\sum_{i=1}^{n} P(A_i \cap A_j)}{\sum_{i=1}^{n} P(A_i)^2} = \frac{\sum_{i=1}^{n} P(A_i)P(A_j) \quad \sum_{i=1}^{n} P(A_i)}{\sum_{i=1}^{n} P(A_i)P(A_j)} = \frac{\sum_{i=1}^{n} P(A_i)P(A_j) - \sum_{i=1}^{n} P(A_i) \cdot \sum_{i=1}^{n} P(A_i)^2}{\sum_{i=1}^{n} P(A_i)P(A_j)}.$$

Now

$$0 \leq \frac{\sum_{i=1}^{n} P(A_i)P(A_j) - P(A_j)}{\left[\sum_{i=1}^{n} P(A_i)\right]^2} \leq \frac{\sum_{i=1}^{n} P(A_i)P(A_j)}{\left[\sum_{i=1}^{n} P(A_i)\right]^2} \to 0 \text{ as } n \to \infty$$

hence the limit in question is satisfied.

(c) $P\left\{\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} P(A_i)}{n}\right| \geq \varepsilon\right\} \leq \frac{\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right]}{\varepsilon^2 \left[\sum_{i=1}^{n} P(A_i)\right]^2}$

Now

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left[\sum_{i=1}^{n} E[X_i]\right]^2$$
$$= \sum_{i=1}^{n} P(A_i) - \left[\sum_{i=1}^{n} P(A_i)\right]^2.$$

The "if" hypothesis implies the desired result.

(c) Let

$$a_n = P\left\{\sum_{i=1}^{n} X_i < \frac{1}{2}\sum_{i=1}^{n} P(A_i)\right\}$$
$$\le P\left\{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} P(A_i) < -\frac{1}{2}\sum_{i=1}^{n} P(A_i)\right\}.$$

Then since $\sum_{i=1}^{\infty} P(A_i) = \infty$ by (b), there is an integer $n_1 > 0$, $n_1 < \infty$ such that $\sum_{i=1}^{n_1} P(A_i) > 0$. By the Borel–Cantelli lemma, with probability 1 we have

$$\sum_{i=1}^{n_1} X_i \ge \frac{1}{2}\sum_{i=1}^{n_1} P(A_i)$$

for only finitely many j, that is,

$$\sum_{i=1}^{n_j} X_i \ge \frac{1}{2}\sum_{i=1}^{n_j} P(A_i)$$

for larger and larger j.

(d) By (c), $\sum_{i=1}^{n} X_i$ is always $\ge$ or $<$ half of $\sum_{i=1}^{n} P(A_i)$ with probability 1, infinitely many A_i occur; thus $P(\limsup_n A_n) = 1$.

Section 4.2

1. Let $A_n = \sum_{i=1}^{n} Y_i$ of $\{Y_n\}$ converges to ∞ in the mean, then

$$\frac{A_n}{n} = \frac{S_n}{n} - \frac{A_{n-1}}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

By the second Borel–Cantelli lemma, $\sum_{n=1}^{\infty} P(|X_n| > \varepsilon) < \infty$, since all X_n have the same distribution. $\sum_{n=1}^{\infty} P(|X| \geq n) < \infty$ so by 6.2.5, $E|X| < \infty$. But then $X_n/n \to 0$ a.s. by 6.2.3.

4. Let $\lambda = (\alpha_1, \ldots, \alpha_k)$. Then $g(\lambda P) = g(\alpha_1 P^{(1)}, \ldots, \alpha_k P^{(k)})$ whenever P permutes cells by many coordinates. Thus $\{g \geq \alpha\}$ is symmetric, and hence $\geq \alpha$ by the Hewitt–Savage zero-one law for probability. Let $\alpha = \inf\{\alpha : P(g \geq \alpha) = 0\}$, then $g \leq \alpha$ a.s.

7. (a) Let π be a permutation of $\{1, \ldots, n\}$ and π^{-1} its inverse. Then $P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_{\pi(1)} \leq x_1, \ldots, X_{\pi(n)} \leq x_n)$. (Note $\{X_{\pi(1)} \leq x_1, \ldots, X_{\pi(n)} \leq x_n\} = \{X_1 \leq x_{\pi^{-1}(1)}, \ldots, X_n \leq x_{\pi^{-1}(n)}\}$.) So all k-dim $F(X = x) = G(x)$. It follows that if ν is Lebesgue measure, $P(X \in A) = \nu(A)$ for every cubic interval $A \subset [0,1]^k$, and hence for every interval $A \subset [0,1]^k$ by continuity. Thus $P(a \leq X \leq b) = \prod_{i=1}^{k} (b_i - a_i)$ as desired.

(b) Fix n and i, and let $A_j = \{x : x_i \in [j/n, (j+1)/n)\}$. Let $\mathcal{A}$ be the collection of finite disjoint unions of the cubic expansion of A_j in n. Then $\mathcal{A}$ is the intersection of $\{X_i = j\}$.

$$P^{\mathcal{A}}(\{X_i = j\}) = P\left\{ \frac{1}{n^{k-1}} \left(\sum_{\substack{m=1 \\ m \neq i}}^{k} 2^{m-1} X_m \right) + \frac{j}{n} \right\} = 1$$

by the strong law of large numbers. If

$$A = \bigcup_{\substack{m=1 \\ m \neq i}}^{k} \bigcup_{j=1}^{2^{m-1}} A_{m,j}$$

it follows that $P^{\mathcal{A}}(A) = P\{X \in A\} = 1$. But $P^{\mathcal{A}}$ is Lebesgue measure on part (a), and the result follows.

(c) We must write

$$\int_0^1 K(x,y)\,dx = \int_0^1 K_1(x)\,dF(x) - \int_0^1 K_1(x)\,dF(x) = E\{K_1(X_1) \mid X_2 = y\} - G(y)$$

and similarly

$$\int_0^1 K(x,y)\,dy = E\{K_2(X_2) \mid X_1 = x\} - G(x)$$

where $K_1(X_1)$ and $K_2(X_2)$ are independent. Since $E(K) = 0$ and $E(K_1(X_1)) = 0$, the result follows.

Section 6.5

5 Recall A_{n+1} is a countable union of sets in $\mathcal{A}_{n+1}$; hence

$$\begin{aligned}\int_{A_{n+1}} X_{n+1}\,dP &= \sum_i \int_{A_{n+1,i}} X_{n+1}\,dP \\ &= \sum_i \lambda(A_{n+1,i}) \qquad \text{by definition of } X_{n+1} \\ &= \lambda(A_{n+1}) \qquad \text{since } \lambda \text{ is countably additive} \\ &= \int_{A_{n+1}} X_n\,dP \qquad \text{by definition of } X_n.\end{aligned}$$

6 $E(Y_n) = \int_{\{y_{n-1}(x)=0\}} y_n(x)\,dx \le \int_{[0,1]^n} y_n(x)\,dx = c_n$.

Now $\mathcal{A}_n = \{[0,Y_1)\times\cdots\times[0,Y_n) : Y_i \in \mathcal{B}[0,1], 0 \le Y_i \le 1\}$. Then $\int_{A_n} X_{n+1}\,dP = E(Y_{n+1}I_{A_n})$ hence

$$\int_{A_n} X_{n+1}\,dP = \int_{\{y_{n-1}(x)>0\}} \frac{y_{n+1}(x)}{y_{n-1}(x)}\,dx_1\cdots dx_n \qquad (*)$$

where $x = (x_1,\ldots,x_{n+1}) \in A_n^c = [0,1]^{n+1} - [0,Y_1)\times\cdots\times[0,Y_n)$ the area of exactly n of the $n+1$ unit squares in $[0,1]^{n+1}$. Now $y_n(x) = \int_0^{x_n} y_{n-1}(x)\,dx_{n-1}$, thus if $x \in B$ and $y_n(x) = 0$, then $y_{n-1}(x) = 0$ except for x_{n-1} in a set of Lebesgue measure 0, so the integration of $y_{n-1}(x)$ with respect to x_{n-1} in $(*)$ will be 0 in each case, i.e.

$$\int_{\{y_{n-1}(x)=0\}\cap[0,Y_1)\times\cdots\times[0,Y_n)} y_{n-1}(x)\,dx = 0$$

Therefore, we right since $x \in A_n$ means

$$\begin{aligned}&\int_{\{y_{n-1}(x)>0\}\cap[0,Y_1)\times\cdots\times[0,Y_n)} \frac{y_{n+1}(x)}{y_{n-1}(x)}\,dx \\ &\quad = \int_{[0,Y_1)\times\cdots\times[0,Y_n)} y_{n+1}(x)\,dx \\ &\quad = \int_{[0,Y_1)\times\cdots\times[0,Y_n)} y_n(x)\,dx \\ &\quad = \int_{[0,Y_1)\times\cdots\times[0,Y_n)} \frac{y_n(x)}{y_{n-1}(x)}\,dx \\ &\quad = E(Y_n I_{A_{n-1}}) = \int_{A_{n-1}} X_n\,dP\end{aligned}$$

proving the super martingale property

6. (a) $A_n = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} E(X_i \mid \mathcal{F}_{i-1})$

$$Y_n = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i \mid \mathcal{F}_{i-1}).$$

So $Y_n - A_n = Z_0$.

(b) $E(A_{n+1} - A_n \mid \mathcal{F}_n) = E(X_{n+1} \mid \mathcal{F}_n) - E(X_{n+1} \mid \mathcal{F}_n) = 0$

(c) $A_{n+1} - A_n = X_n - E(X_n \mid \mathcal{F}_{n-1}) \geq 0$ a.s.

Section 6.4

1. (a) $E(X_{n+1} \mid X_n = i) = \mu_{i+1} + \mu_{i-1}$ $(i \geq 1)$, $E(X_{n+1} \mid X_n = 0) = \mu_1$.

$E(X_{n+1} \mid \mathcal{F}_n) = X_{n-1} + \mu_{i-1} + \mu_{i+1}$ $(i \geq 1)$, $E(X_{n+1} \mid \mathcal{F}_n) = X_{n-1} + \mu_1$ $(i = 0)$.

The result is the martingale property. Since all $\mu_i \geq 0$ for all i either $X_n(\omega) \to \infty$ for all ω or else some j, $X_n(\omega) = j$ for $n \geq i$. X_n converges a.s. to ∞.

(b) $E(X_1) = 2\mu_2 + \mu_0$, $E(X_2) = 2\mu_3 + \mu_1$, $E(X_3) = 2\mu_4 + \mu_2$.

$E(X_n) = \mu_{n+2} + \mu_n - (1 - \mu_1)^{-1}\mu_0\mu_2 + \mu_{n-1} - \mu_{n+1}(1 - \mu_1) = 2\mu_{n+1}(\mu_{n+2} - \mu_1)$.

and so on. Thus

$$\lim_{n \to \infty} E(X_n) = \left[\prod_{i=1}^{\infty} (1 - \mu_i)\right]^{-1} \sum_{i=1}^{\infty} \mu_i \leq \infty.$$

The infinite product is greater than 0 since $\sum \mu_i < \infty$, hence $E(X_n) < \infty$.

2. By definition of the process $X_n(X_{n-1}, Y_n, \ldots, X_0) = X(X_{n+1}, X_n)$. Therefore the matrix driving these processes is the matrix of transition probs.

$$E\begin{pmatrix} X_{n+1} \\ X_n \end{pmatrix} \mid X_n = \begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} p_{ij} & p_{i+1,j} \\ p_{ij-1} & p_{i+1,j-1} \end{pmatrix} + \begin{pmatrix} i \\ j \end{pmatrix} - \begin{pmatrix} i \\ j \end{pmatrix}$$

Thus $E(X_{n+1} \mid X_n) = X_n$. Since $|X_n| \le 1$ for all n, $Y_n \to X_\infty$ a.s. and in $\mathcal{L}^1$, by the martingale convergence theorem, $E(X_n) \to \lim_n E(X_n) = E(X_\infty)$.

Section 6.3

1. Since $|f_n + g_n - (f + g)|^p \le 2^{p-1}(|f_n - f|^p + |g_n - g|^p)$, $p \ge 1$, and $|f_n|^p \vee |g_n - g|^p \le |f_n - f|^p + |f|^p$, $p \ge 1$, by 6.3.3 it suffices to show that the $f_n - f$ are uniformly integrable. Now

$$\int_A |f_n - f|^p \, d\mu \to 0 \quad \text{as} \quad \mu(A) \to 0$$

for any fixed n, and

$$\int_A |f_n - f|^p \, d\mu \le \int_{A \cap [0,n]} |f_n - f|^p \, d\mu + \int_{[n,\infty)} |f_n - f|^p \, d\mu \to 0 \quad \text{as} \quad n \to \infty$$

by the $\mathcal{L}^p$ convergence. It follows that the integrals of $|f_n - f|^p$ are uniformly small. For f continuous and uniformly vanished, the result follows from 6.3.5.

Section 6.4

4. (a) If the Y_n are uniformly integrable, $E(Y_n) \to E(Y_\infty)$ by 6.3.3. Conversely, assume $E(Y_n) \to E(Y_\infty)$. If ε stands for $\min\{\varepsilon, \varepsilon/2\}$, above $|Y_n - \lambda_n| - |Y_\infty - \lambda_\infty| \le |(Y_n - \lambda_n) - (Y_\infty - \lambda_\infty)|$ and $Y_n + Y_\infty - (Y_n \vee Y_\infty) = (Y_n \wedge Y_\infty) \ge 0$, it follows that $E(|Y_n - \lambda_n|) - E(|Y_\infty - \lambda_\infty|) \le E(Y_n + Y_\infty - (Y_n \vee Y_\infty)) \to 0$. Thus $\lim_n E(|Y_n - \lambda_n|) = E(|Y_\infty - \lambda_\infty|)$, and by the dominated convergence theorem, $E(|Y_n - \lambda_n| \wedge \varepsilon) \to E(|Y_\infty - \lambda_\infty| \wedge \varepsilon)$. Hence $\limsup_n E(|Y_n - \lambda_n|) \le \lim_n E(|Y_n - \lambda_n| \wedge \varepsilon) + \varepsilon = E(|Y_\infty - \lambda_\infty| \wedge \varepsilon) + \varepsilon \le \varepsilon$. Thus $Y_n - \lambda_n \to 0$ in $\mathcal{L}^1$, and it follows that the Y_n are uniformly integrable. (See Problem 6.3.5, in general, for a converse when $|f_n|$ implies uniform integrability of $|f_n|^p$.)

   (b) If $A = \mathcal{A}_\infty$, $\pi = \pi_\infty$, then $\int_A X_n \to \int_A X_\infty$. For $n \to \infty$ we apply Fatou's lemma,

$$\int_A X_\infty = \int_A \liminf_n X_n \le \liminf_n \int_A X_n \le \int_A X_\infty$$

   (c) By Fatou's lemma, $E(X_{\infty,\pi}) = E(\liminf_\pi X_n) \le \liminf_\pi E(X_n) = 0$.

Section 6.7

1. $|X_T| = X_0 + X_1 - 2X_0^+ = X_T$, hence $E(|X_T|) \le 2E(X_0^+) + E(X_1^+) + E(X_T^+)$ by 6.7.1. But $X_T^+, \ldots, X_0^+$ is a submartingale by 6.2.6(a); hence $E(X_T^+) \le E(X_0^+)$ by 6.7.1, as desired.

1. (a) Define T as indicated. By 6.7.3, $|X_n, X_{n+1}|$ is a submartingale; hence

$$E(X_{n+1}) \geq E(X_n) = \int_{[\max(X_n, X_{n+1}) \geq \lambda]} X_n \, dP \geq \int_{[\max(X_n, X_{n+1}) \geq \lambda]} X_{n+1} \, dP$$

$$\geq \lambda P[\max(X_n, X_{n+1}) \geq \lambda] = \int_{[T \leq n+1]} Y_T \, dP$$

and the result follows.

(b) By 6.7.3, $|X_n, X_{n+1}|$ is a supermartingale; hence

$$\lambda P[T = n+1] = \lambda(F_{n+1} - F_n) = \int_{[\max(X_n, X_{n+1}) < \lambda]} X_n \, dP - \int_{[\max(X_n, X_{n+1}) < \lambda]} X_{n+1} \, dP$$

$$\geq \lambda P[\max(X_n, X_{n+1}) \geq \lambda] - \int_{[\max(X_n, X_{n+1}) \geq \lambda]} X_{n+1} \, dP.$$

Since $\max(X_n, X_{n+1}) \geq |X_{n+1}| = X_{n+1}$, the result follows.

(c) Since

$$\left\{ \max_{1 \leq k \leq n} X_k \geq \lambda \right\} \downarrow \left\{ \sup_k X_k \geq \lambda \right\} \quad \text{as} \quad n \to \infty,$$

the result follows from (a) and (b). Note also that the same argument holds with $\{\sup_k X_k \geq \lambda\}$ replaced by $\{\sup_k X_k > \lambda\}$; this follows easily from

$$\left\{ \sup_k X_k > \lambda \right\} = \bigcup_j \left\{ \sup_k X_k \geq \lambda + \frac{1}{j} \right\}.$$

5. (a) Since $X_n = \sum_{k=1}^{n} (Z_k - \mu)$ and $E(Z_k - \mu) = 0$, $|X_n|$ is a submartingale. By 6.7.3, $E|X_{n+1}| - E|X_n| = E|X_{n+1}| - E|X_{T \wedge n}| \geq 0$. Since $E|X_{T \wedge n}| \leq \lambda$, $\lim_n E|X_n| = \lim_n E|X_{T \wedge n}|$ exists. Since $P[T < \infty] = 1$, $X_{T \wedge n} \to X_T$ a.e. P as $n \to \infty$, and by the monotone convergence theorem, $E|X_T| = \lim E|X_{T \wedge n}|$.

(b) We write $X_{n+1} = \sum_{k=1}^{T} Z_k - \sum_{k=1}^{n} Z_k = Z_{n+1} \cdot I_{[T \geq n+1]}$. By (a), $E(Z_{n+1}) = E(Z^+)P[T \geq n+1] - E(Z^-)P[T \geq n+1] = E(Z^+) - E(Z^-)$. Since $E(Z^+)$ is finite, so are $E(X_n^+)$ and $E(X_n^-)$; hence $E(X_{n+1}) < \infty$ and

$$E(X_{n+1}) = E(X_n^+) - E(X_n^-) + E(Z^+) - E(Z^-) \geq E(X_n^+) - E(X_n^-) = E(X_n).$$

(c) To prove (c), observe first that if all $Y_n \geq 0$, then

$$E(X_T) = \sum_{n=1}^{\infty} E(X_n \cdot I_{[T=n]})$$

$$= \sum_{n=1}^{\infty} E(Y_n) P[T \geq n] \quad \text{by independence}$$

$$= n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

Thus D_n is given just as above.

Section 6.6

8. (a) $\{X_n\}$ is irreducible and is a finite birth and death chain with

$$P\{X_{n+1} = X_n + 1 \mid X_n = i\} = \frac{\lambda}{\lambda + i\mu} \quad \text{for } i = 1, 2, \ldots, N-1,$$

so that $X_{n+1} = X_n \pm 1$ eventually.

(b) Since $X_0 \geq 0$, $E[X_{n+1} \mid X_n] \geq 2$, hence by the monotone convergence theorem,

$$E[X_{T-1}] = \lim_{s \to 1} \sum_{k=0}^{\infty} s^k P\{X_{T-1} = k\} \geq \lim_{s \to 1} \sum_{k=0}^{1} s^k P\{X_{T-1} = k\}$$

Given $X_{T-1} = k$, $k \geq 1$, we have $X_T = X_{T-1} - 1$ or $X_{T-1} + 1$. But

$$\lim_{s \to 1} \sum_{k=0}^{1} s^k P\{X_{T-1} = k\} = \lim_{s \to 1} \sum_{k=0}^{1} s^k P\{X_T = k\}$$

by the intermediate property.

(c) T is the time at which the betting stops. In this case, T is also the time of going broke. By (a), T is also finite, and we must follow.

(d) Realistically, there is a time at which one can leave the game, whether one is losing or winning. Also, a casino has a limit on how much it will allow a player to win. Finally, we have not been subject to inflation.

9. The probability of reaching $r+1$ before 0 is $x/(r+1)$, and

$$E[Z_1] = \frac{r}{r+1}(r+1) + \frac{1}{r-1}(0) - 1 = E[Z_0]$$

(See Ash, 1970, 6.2 for details.)

10. (a) If $\sum_{i=1}^{\infty} p_i = \infty$ and $\sum_{i=1}^{\infty} q_i < \infty$, then $X_n \to \infty$.

(b) If $\sum_{i=1}^{\infty} p_i < \infty$ and $\sum_{i=1}^{\infty} q_i = \infty$, then $X_n \to -\infty$.

Hint: $\{X_n\}$ is a martingale by Problem 9, hence by (a) of 8, X_n converges a.e. to a finite limit or else jump X_n is such that $X_n \to \infty$ or $-\infty$. In case (a), if $X_n \to \infty$ and in case (b) if $X_n \to -\infty$ we have a contradiction in the sets

$$\left\{\sum_{i=1}^{\infty} Y_i < \infty, \ \sum_{i=1}^{\infty} q_i < \infty\right\} \quad \text{and} \quad \left\{\sum_{i=1}^{\infty} Y_i = \infty, \ \sum_{i=1}^{\infty} q_i = \infty\right\}$$

now probability one. Note that $|Y_n - q_n| \leq 2$ so 0.0.1 actually applies.

Chapter 7

Section 7.1

1. (a) $\mathcal{L}[f(t)](s) = \int_0^\infty e^{-st} \cosh(at)\,dt = e^{st}$ for some s means $e^{(s-a)t} = \frac{1}{2}e^{(s+a)t}\,dt$ or $1 = \frac{1}{2}e^{(s+a)t}\,dt$. Take real parts to obtain $\frac{1}{2}e^{(s+a)t} = \cos(\theta)$ and $\frac{1}{2}e^{(s-a)t} = 1$. Since the integrand is convergent, we have $\cos(\theta) = 1$ and $\frac{1}{2}e^{(s-a)t} = 1$. But $\cos(\theta) = 1$ if $\theta = 2\pi n$, n an integer. Thus $\theta = 2\pi n$. This is infinitely distinct. The inverse is proved by reversing the argument.

   (b) By part (a):

$$\mathcal{L}[f(t)](s) = \frac{1}{2}\left[\frac{1}{s-a} + \frac{1}{s+a}\right] \qquad \text{for some integer } n$$

$$= \frac{s}{s^2-a^2} + 2\pi n i\left[\frac{1}{s-a} - \frac{1}{s+a}\right] \qquad \text{for some integer } n \neq 0$$

for appropriate s and a. If $f(t)$ is nonzero, then $\mathcal{L}[f(t)](s)$ is unbounded, and $\frac{s}{s^2-a^2} + 2\pi n i\left[\frac{1}{s-a} - \frac{1}{s+a}\right]$ for an integer n must have at least two points in common, and this implies that $2\pi n i$ and $2\pi n i$ are not equal. This is a contradiction.

5. (a) By 7.1 Step A, f has continuous derivatives on $\mathbb{R}$ and $f^{(k)}(0) = 0$ for $k = 0, 1, \ldots, n$. Now, $f(t) = 0$, where I is an interval of $\mathbb{R}$ containing 0 and $\mathbb{R}$ has continuous derivatives on I, then $f \equiv 0$.

$$\sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}t^k + \frac{1}{n!}\int_0^t (t-\tau)^n f^{(n+1)}(\tau)\,d\tau = f(t).$$

(This is an exercise in calculus; see Apostol, 1969, p. 173 for details.) Add and subtract

$$\frac{f^{(n+1)}(0)}{n!}t^n = \frac{1}{n!}\int_0^t (t-\tau)^n f^{(n+1)}(\tau)\,d\tau$$

from the above equation to obtain

$$f(t) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}t^k + R_n(t)$$

where

$$R_n(t) = \frac{1}{n!}\int_0^t [(t-\tau)^n - t^n] f^{(n+1)}(\tau)\,d\tau$$

Since $M_n(s)\to e^{\sigma^2 s^2/2}$ as $n\to\infty$ by the dominated convergence theorem. The result follows.

(b) By Problem 6.7(c), if $\mathbf{X}$ has a multivariate normal distribution, then it can be seen that

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \sum_{k=0}^{n-1} \frac{i^k \varphi^{(k)}(0)}{k!} t^k + i^n \int_0^t \frac{(t-s)^{n-1}}{(n-1)!} \varphi^{(n)}(s)\, ds.$$

Now $\varphi^{(n)}(s) = \int_{\mathbb{R}^n} (ix)^n e^{isx}\, dF(x)$ by Problem 6.7(c) because $\int |x|^n\, dF(x) \le E|X|^n$; the result follows.

(c) $\int \left(\frac{e^{itx}-1}{x}\right)^2 dF(x) = \int_{-\infty}^{\infty} \left(\frac{\cos tx - 1}{x}\right)^2 dF(x) + \int_{-\infty}^{\infty} \left(\frac{\sin tx}{x}\right)^2 dF(x)$

$$= \int_{-\infty}^{\infty} \left(\frac{\cos tx - 1}{x^2}\right)^2 x^2\, dF(x) + \int_{-\infty}^{\infty} \left(\frac{\sin tx}{x}\right)^2 dF(x)$$

(d) By L'Hôpital's rule,

$$\lim_{t\to 0} \frac{1}{t^2} \int \left(\frac{e^{itx}-1}{x}\right)^2 dF(x) = \lim_{t\to 0} \frac{\varphi'(t+\delta t) - \varphi'(t) - \varphi'(\delta t)}{\delta t}$$

$$= \lim_{t\to 0} \frac{\varphi'(t+\delta t) - \varphi'(t) - \varphi'(0)}{\delta t} - \frac{\varphi'(\delta t) - \varphi'(0)}{\delta t} = \varphi''(0) - \varphi''(0).$$

Thus

$$\varphi''(0) = \lim_{t\to 0} \int_{-\infty}^{\infty} \left(\frac{\sin tx}{tx}\right)^2 x^2\, dF(x)$$

$$\ge \int_{-\infty}^{\infty} \liminf_{t\to 0} \left(\frac{\sin tx}{tx}\right)^2 x^2\, dF(x) \qquad \text{by Fatou's Lemma}$$

$$= \int_{-\infty}^{\infty} x^2\, dF(x).$$

(e) Assume the result holds up to the integer n and assume $\int |x|^{n+1}\, dF(x)$ exists and is finite. Then $\int_{-\infty}^{\infty} |x|^n\, dF(x) < \infty$ so by (b) we have

$$\varphi^{(n)}(t) = \int_{-\infty}^{\infty} (ix)^n e^{itx}\, dF(x).$$

or

$$\frac{1}{n!}\int_{-\infty}^{\infty} x^n \, dG(x) = \int_{-\infty}^{\infty} x^n \, dF(x),$$

where

$$G(x) = \int_{-\infty}^{x} |t|^n \, dF(t).$$

Since G is a bounded distribution function with characteristic function $\int_{-\infty}^{\infty} e^{itx} \, dG(x)$, part (b) shows that $\int_{-\infty}^{\infty} x^n \, dG(x) < \infty$, and so $\int_{-\infty}^{\infty} |x|^n \, dF(x) < \infty$. The result follows by induction.

Section 9.2

4. (a) Choose $a > 0$ such that $I = \int_{-a}^{a} g(x) \, dx \neq 0$. (This is possible since $\int_{-\infty}^{\infty} g(x) \, dx = 1$.) Since g is continuous at $x = 0$, there exists $\delta > 0$ such that $|g(x) - g(0)| < |I|/2a$ whenever $|x| < \delta$. Then, for $0 < \varepsilon < \delta$,

$$\frac{\int_{-a}^{a} \exp(itx/\varepsilon) g(x) \, dx}{\int_{-a}^{a} \exp(itx/\varepsilon) g(0) \, dx} = \frac{\int_{-a}^{a} \exp(itx/\varepsilon)\{g(x) - g(0)\} \, dx}{\int_{-a}^{a} \exp(itx/\varepsilon) g(0) \, dx} + 1 \to \frac{g(0) - g(0)}{I} + 1 = 1$$

as $\varepsilon \to 0$.

(b) This is immediate from (a).

5. (a) Let $F_n(x) = 1$ if $x \geq n$, $= 0$ if $x < n$, so F_n is the distribution function of the random variable $X_n = n$ (a.s.). Then $\phi_n(t) = 1$ for all n. Then $\phi_n(t) \to \phi_0(t)$ for all t, so that $|\phi_n(t)| \to |\phi_0(t)|$, but $\phi_0(t) = 0$ if $t \neq 0$, so ϕ_0 does not correspond to any random variable or F_0.

(b) F_n is uniform, but $F_n(x) \to 0$ for all x, so F_n corresponds to $X_n = n$ (a.s.). If n is odd, let $F_n(x) = 1$ if $x \geq -n$, $= 0$ if $x < -n$; if n is even, let $F_n(x) = 0$ if $x \geq n$, $= 1$ if $x < n$. Then $F_n(x, \infty) = F_n(-\infty, b]$ for all n, $b \in \mathbb{R}$, but $F_n(-\infty, \infty) = 1 \not\to 0 = F_n(-\infty, \infty)$. Furthermore, $\lim_{n \to \infty} \phi_n(t)$ does not exist for any $t \neq 0$.

6. If F_n converges vaguely to F, then $\{F_n\}$ is relatively compact, and hence tight, by (9.14). Thus, given $\varepsilon > 0$, let a and b be finite continuity points of F such that $F_n(\mathbb{R} - (a, b]) < \varepsilon$ for all n. Then

$$\limsup_{n \to \infty} F_n(\mathbb{R}) \leq \varepsilon + \limsup_{n \to \infty} F_n(a, b]$$
$$= \varepsilon + F(a, b]$$
$$\leq \varepsilon + F(\mathbb{R}).$$

But $F_n(\theta) \geq F_n(x, 0]$; hence

$$\liminf_{n\to\infty} F_n(\theta) \geq F(x, 0].$$

Since x is a boundary and θ may be taken arbitrarily large and x arbitrarily small, we have $\liminf F_n(x) \geq F(x)$. A similar argument shows that $F_n(x, \theta] \to F(x, \infty]$ and $F_n[a, \infty] \to F[a, \infty]$ if a and b are continuity points of F. Therefore F_n converges weakly to F.

4. (a) $E(X_{n+1}|\mathcal{F}_n) = E(X_n \theta_{n+1}|\mathcal{F}_n) \exp(\alpha S_{n+1} - \gamma S_n)$

$$= X_n E(\theta_{n+1}) \exp(\alpha S_n - \gamma S_n) E(\exp(\alpha Z_{n+1}))$$

since X_n is $\mathcal{F}_n$ measurable and the θ_n are independent

$$= X_n.$$

(b) By hypothesis, $\bigcap_{n=1}^{\infty} I_n = \Delta_X$ with only at most countably many bounded intervals. Thus Δ_X is a countable open interval consisting of all x such that $\lim_{n\to\infty} S_n(x) = 0$. Since $\{S_n(x)\}$ is a normal sequence, $\{X_n, \mathcal{F}_n\}$ is a martingale and hence converges for $x \in \Delta_X$:

$$X_n(x) = \left[\prod_{i=1}^{n} \theta_i(x)\right] \exp\left[\alpha \sum_{i=1}^{n} Z_i(x) - \gamma n\right],$$

and the result follows.

(c) Let C be the subset $\{(\omega, x) : x \in \Omega, \omega \in \Omega\}$ such that

$$\exp\left[\alpha \sum_{i=1}^{n} Z_i(\omega) - \gamma n\right]$$

fails to converge. By (b), $\{\omega : (\omega, x) \in C\}$ has probability 0 for each $x \in \Omega$, so by Fubini's theorem, Section 2.6, $\{x : (\omega, x) \in C\}$ has Lebesgue measure 0 for almost every ω.

(d) Convergence a.s. and in probability are equivalent in probability. Since a subsequence of a convergent sequence converges in probability to the same limit, convergence in probability implies convergence in distribution by 7.1.1. By parts (b) and (c), convergence in distribution implies convergence a.s.

Section 2.3

3. Let the X_{nk}'s be as in the hint. Then

$$\begin{aligned}\max_{0\le t\le 1}\left|\hat{f}_n(t)-\mathbb{E}\hat{f}_n(t)\right| &= \max_{0\le t\le 1}\left|\int_{-\infty}^{\infty} e^{itx}\,d(F_n(x)-F(x))\right| \\ &\le \max_{0\le t\le 1}\int_{-\infty}^{\infty}|e^{itx}-1|\,d|F_n(x)-F(x)| \\ &\le \sup_{|x|\le a}|e^{itx}-1|\int_{|x|\le a}d|F_n(x)-F(x)| \\ &\quad + \int_{|x|>a}2\,d|F_n(x)-F(x)|.\end{aligned}$$

Now $|e^{ix}-1|\le|x|$; hence

$$\begin{aligned}\sup_{|x|\le a}|e^{itx}-1| &\le \max_{1\le k\le n}\int_{X_{nk}}^{X_{nk}+h}|x-X_{nk}|\,dF_{X_{nk}}(x) \\ &= \max_{1\le k\le n}\int_{X_{nk}}^{X_{nk}+h}|x-X_{nk}|\,dF_{X_{nk}}(x) \\ &\le |t|h=h\max_{1\le k\le n}P\{|X_{nk}-\mu_k|\le h\}.\end{aligned}$$

The second term approaches 0 as $n\to\infty$ for any $a>0$ by the uniform asymptotic negligibility of the X_{nk}'s. Thus

$$\max_{0\le t\le 1}|\hat{f}_n(t)-\mathbb{E}\hat{f}_n(t)|\to 0 \quad \text{as} \quad n\to\infty$$

uniformly for t in a finite interval.

Conversely, suppose

$$\max_{0\le t\le 1}|\hat{f}_n(t)-\mathbb{E}\hat{f}_n(t)|\to 0.$$

By the triangle inequality we have

$$\begin{aligned}\max_{1\le k\le n}P\{|X_{nk}-\mu_k|\le h\} &= \max_{1\le k\le n}\int_{|x-\mu_k|\le h}dF_{X_{nk}}(x) \\ &\le \max_{1\le k\le n}2a\int_0^{1/h}|\hat{f}_{X_{nk}}(t)|\,dt \\ &\le 2a\int_0^{1/h}\max_{1\le k\le n}|\hat{f}_{X_{nk}}(t)|\,dt\end{aligned}$$

since $|\hat{f}_{X_{nk}}(t)|=|\mathbb{E}\cos t(X_{nk}-\mu_k)|\le 1$.

This suggests approaches like those in the hypothesis and the classical convergence theorem, and the result follows.

4. (a) We let $X_0 = 1$ and $\sigma_0 = \sqrt{n}$. Now, for a given $\varepsilon > 0$, we let n be large enough so that $n\sqrt{n} > 1$. Then

$$P\left\{\left|\frac{X_n}{\sigma_n}\right| \geq \varepsilon\right\} = P\{|X_n| \geq \varepsilon\sqrt{n}\}$$

$$\begin{cases} 0 & \text{if } \varepsilon > n\sqrt{n} \\ \frac{1}{n^2}\left(1 - \frac{1}{n}\right) & \text{if } \varepsilon \leq n\sqrt{n} \end{cases}$$

Thus

$$\lim_{n\to\infty} P\left\{\left|\frac{X_n}{\sigma_n}\right| \geq \varepsilon\right\} = \frac{1}{n^2}\left(1 - \frac{1}{n}\right) \to 0$$

(b) $\frac{1}{n}\sum_{i=1}^{n} \int_{|x-\mu|>\varepsilon n} (x-\mu)^2 \, dF(x) = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2; |X_i| > \varepsilon n]$.

As in (a) we let n be large enough so that $\varepsilon\sqrt{n} > 1$. We obtain

$$\frac{1}{n}\sum_{i=\sqrt{n}}^{n} E[X_i^2; |X_i| > \varepsilon n] = \frac{1}{n}(n - \sqrt{n})\sqrt{n}\left(1 - \frac{1}{n}\right) = 1 - \frac{1}{\sqrt{n}} \to 1$$

This time, independence cannot help because $\varepsilon_n \to 0$. Now

$$\text{Var}\, X_n = E[X_n^2; |X_n| \leq \varepsilon n]$$

$$= E[X_n^2] = 1 \qquad \text{if } k \leq \varepsilon\sqrt{n}$$

$$P\{|X_n| = 1\} = \frac{1}{n} \qquad \text{if } k > \varepsilon\sqrt{n}.$$

Thus

$$s_n^2 = \sum_{k=1}^{n} \text{Var}\, X_k = [\varepsilon\sqrt{n}] + \frac{1}{n}(n - [\varepsilon\sqrt{n}]) \sim \varepsilon\sqrt{n}.$$

The Lindeberg sum for S_n/σ_n is

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\sum_{k=1}^{n} E\left[X_k^2 I_{\{|X_k|>\varepsilon\sigma_n\}}\right] \\
&\quad \le \frac{1}{\sigma_n^2}\sum_{k=1}^{n} k^2 P\{|X_1| > \varepsilon\sigma_n\} \\
&\quad = \frac{1}{n}\left(\sqrt{n}\,\varepsilon\sigma_n\right)^{-2}\left(n^2 P\{|X_1| > \varepsilon\sigma_n\}\right) \to 0,
\end{aligned}$$

since $\sigma_n^2/n \to \sigma^2 > 0$ and $n^2 P\{|X_1| > \varepsilon\sigma_n\} \to 0$ because

$$\begin{aligned}
n^2 P\{|X_1| > \varepsilon\sigma_n\} &\le n^2 P\{|X_1| \ge \varepsilon\sigma_n/2\} \\
&\le \sum_{k=\varepsilon\sigma_n/2}^{\infty} k^2 P\{k \le |X_1| < k+1\} \\
&\to 0 \quad \text{as} \quad n \to \infty
\end{aligned}$$

since

$$P\{|X_1| = k\} = \frac{1}{k^2}\left(1 - \frac{1}{k}\right) \quad \text{and} \quad \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty.$$

5.8: If $Y_n \xrightarrow{d} Y$ and $a_n \to 1$, then $a_n Y_n \xrightarrow{d} Y$, for if Δ_n is the characteristic function of Y_n and Δ is the characteristic function of Y, we have $\Delta_n(t) \to \Delta(t)$. Now $\varphi(t)\Delta_n(t) - \Delta(t) = \Delta_n(t)[\varphi(t) - 1] + \Delta(t) - \Delta_n(t)$, where $\varphi(t) \to 1$, so that $\varphi(t)\Delta_n(t) \xrightarrow{u} \Delta(t)$ (5.1.1) by the Arzelà theorem. If $Y_n \xrightarrow{p} Y$ and $P\{Y_n \neq Y_n'\} \to 0$, then $Y_n' \xrightarrow{d} Y$ because

$$\begin{aligned}
F_{Y_n'}(x) &= P\{Y_n' \le x, Y_n' \le x\} + P\{Y_n' \le x, Y_n' > x\} \\
&\le F_{Y_n}(x) + P\{Y_n' \neq Y_n\}.
\end{aligned}$$

Thus by (5.1.2) $\sqrt{n}\,\Phi_n(t) \xrightarrow{d}$ normal (0, 1). But by (5.1.1) and (5.1.3), $S_n/\sigma_n = S_n/\sqrt{n}\,\sigma \xrightarrow{d}$ normal (0, 1).

Section 7.4

1. (a) $P[\text{negative}] = P[X < 0] = P[Z < \frac{0-\mu}{\sigma}]$

$$= \int_{-\infty}^{-\mu/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y+\mu/\sigma)^2\right) dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\mu/\sigma}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy$$

$$= 1 - \frac{1}{\sqrt{2\pi}} \int_{0}^{\mu/\sigma} \exp\left(-\frac{1}{2}y^2\right) dy$$

$$= \frac{1}{\sqrt{2\pi}} \left[\int_{0}^{\infty} - \int_{0}^{\mu/\sigma}\right] \exp\left(-\frac{1}{2}y^2\right) dy = \alpha(\mu),$$

where $\alpha(\mu) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\mu/\sigma} \exp\left(-\frac{1}{2}y^2\right) dy \to 0$ as $\mu \to \infty$ since

$$\int_{0}^{\mu/\sigma} \exp\left(-\frac{1}{2}y^2\right) dy \le \int_{0}^{\mu/\sigma} y^{-2} dy = \frac{\sigma}{\mu}, \quad \mu > 0.$$

The result is now true for $\mu > 0$.

(b) Let $I = \int_{0}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy$. Then (see ex. 6.3) of the proof of Theorem 7.2.1),

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\text{var}(Y_1) = \frac{1}{n}\left[\int_{0}^{\infty} y^2 \exp\left(-\frac{1}{2}y^2\right) dy - \left(\int_{0}^{\infty} y \exp\left(-\frac{1}{2}y^2\right) dy\right)^2\right] = \frac{1}{n}\left[\frac{\pi}{2} - 1\right] = \frac{\pi - 2}{2n}.$$

Hence $a(n) = \sqrt{n}$.

(c) We have

$$\text{var}(Y) = \int_{0}^{\infty} |y|^{2\gamma/(2-\gamma)} \exp\left(-\frac{1}{2}|y|^{\gamma}\right) dy = \int_{0}^{\infty} y^{2\gamma/(2-\gamma)} \exp\left(-\frac{1}{2}y^{\gamma}\right) dy;$$

hence

$$\text{var}(\bar{Y}_n) = n^{-1}\left(-|y|^{2\gamma/(2-\gamma)} \int_{0}^{\infty} y^{\gamma} \exp\left(-\frac{1}{2}y^{\gamma}\right) dy\right).$$

This is of the form $n^{-\alpha}\beta$ with $\alpha = 1$, $0 < \beta < \infty$ if $0 < \gamma < 2$.

Section 7.5

1. A logarithm of the odds is given by

$$\ln\left[\frac{L(\theta_1)}{L(\theta_0)}\right] = \ln\left[\frac{g(\bar{x})}{h(\bar{x})}\right] = \ln\left[\sum_{i=1}^{n} x_i^2\right] - \ln n.$$

Thus

$$M_n(t) = \lim_{n\to\infty} \exp\left\{ \sum_{j=1}^{n} \left[\frac{it}{n} + \frac{1}{2}\left(\frac{it}{n}\right)^2 + o\left(\frac{t^2}{n^2}\right) \right] \right\}.$$

But

$$\exp\left\{ \sum_{j=1}^{n} \left[\frac{it}{n} + \frac{1}{2}\left(\frac{it}{n}\right)^2 + o\left(\frac{t^2}{n^2}\right) \right] \right\}$$

is of the Poisson type (see 7.5.3(a)) with $\lambda = \frac{1}{2}t^2$, $a = 1/n$, and the result follows from 7.5.7.

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{x} dF(y)\, dF(x)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{x} I_{(-\infty,\, x]}(y)\, dF(y)\, dF(x)$$

$$= \int_{-\infty}^{\infty} F_X(x)\, dF_Y(x) = E[F_X(Y)],$$

$$\int_{-\infty}^{\infty} F(x)\, dF(x),$$

$$= \int_{-\infty}^{\infty} F(x)\, dF(x) \quad \text{by a symmetrical argument.}$$

In the case where X has a density,

$$P(X \le Y) = \int_{-\infty}^{\infty} F_X(x)\, dF_Y(x) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} f(x)\, dy \right) f(x)\, dx.$$

Let $x = u - v$ in the inner integral

$$P(X \le Y) = \int_{-\infty}^{\infty} \left[\int_{v}^{\infty} f(u)\, du \right] f(v)\, dv$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{u} f(u)\, dv \right] f(u)\, du$$

and the result follows.

Section 7.6

1 By the strong law of large numbers, $F_n(x)$ converges to $F(x)$ ... and $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$ ...

(This holds for each fixed $x \in \mathbb{R}$, and the exceptional set of measure 0 depends on x.)

Assume now $\lambda_1(x) > 0 > \lambda_2(x)$ and $\lambda_2(x) < \lambda_1(x)$ for $x \in A$, where $P(A) = 0$. Let S be a countable dense subset of $\mathbb{R}$ containing all discontinuities of F. If $x \in A$ then $F_x(y) = \lim_{z \in S,\, z \to x} F(z,y)$ is continuous. If $x \in A$ then $F_x(x,y) > 0$ and $F_x(x^-,y) < 0$ for all $y \in S$. Furthermore, $F_x(x,y) = 1$, $F_x(x^-,y) = -1$, $F_x(x,y) = \lambda_1(x) F_x(x,y)$, $F_x(x^-,y) = \lambda_2(x) F_x(x^-,y)$, and $F_x(x,y) + F_x(x^-,y)$ is monotone in y.

Chapter 8

Section 8.2

1. If $T^{-1}A \subset A$, then $A^c \subset T^{-1}A^c$, and conversely, also $A = T^{-1}A = T^{-1}A^c$. This shows that the two definitions of indecomposability are equivalent.

(2) (i) $\Rightarrow$ (ii): Let A be wandering, and let $B = \bigcup_{n=1}^{\infty} T^{-n}A$. Then $T^{-1}B = \bigcup_{n=1}^{\infty} T^{-n}A \subset B$, and so $P(B) = P(T^{-1}B) = P(B \cap T^{-1}B)$. But $B \cap T^{-1}B = A$, since the $T^{-n}A$ are disjoint; hence $P(A) = 0$.

(ii) $\Rightarrow$ (i): Let $A^{(n)} = A \cap T^{-1}A \cap \cdots \cap T^{-n+1}A$. If $C = A - A^{(n)}$, then $T^{-n}C = [x \in A : T^n x \notin A]$. Thus $C \cap T^{-n}C = \emptyset$. Thus the $T^{-n}C$ are disjoint, so C is wandering. By (ii), $P(C) = 0$, and therefore T is measure-preserving.

(iii) $\Rightarrow$ (iv): Let $T^{-1}A \subset A$. Then by induction, $T^{-n}A \subset A$, $n \geq 1$. Thus $T^{-n}A = \bigcap_{k=1}^{n} T^{-k}A$. Now

$$A - T^{-1}A = A - \bigcap_{n=1}^{\infty} T^{-n}A = A - A^{(\infty)},$$

which has measure 0 by (iii). Thus $P(A) = T^{-1}A$, proving (iv).

(iv) $\Rightarrow$ (iii): Obvious.

(v) $\Rightarrow$ (i): Let

$$A^{(\infty)} = A \cap \bigcap_{n=1}^{\infty} T^{-n}A = [x \in A : T^n x \in A \text{ for infinitely many } n \geq 1].$$

If $A \neq A^{(\infty)}$, let $B = \bigcup_{n=1}^{\infty} T^{-n}A$. Then $T^{-1}B \subset B$, hence $T^{-1}(A - B) = A - B$, and $P(A - B) = 0$. Similarly, $T^{-1}(A^c - B^c) = A^c - B^c$, hence

$$P(A^c - B^c) = P(A - B) = 0.$$

But

$$A \setminus B = T^{-1}A \setminus B$$
$$= \bigcup_{k=0}^{\infty} T^{-k}A - \bigcup_{k=1}^{\infty} T^{-k}A$$
$$= \{x : T^k x \text{ enters } A \text{ for the last time at } k = 0\}.$$

Since

$$A - A^* = A \cap \bigcup_{n=1}^{\infty} \{x : T^n x \in A \text{ for the last time at } n = k\},$$

we have

$$\mu(A - A^*) \le \sum_{k=1}^{\infty} \mu(T^{-k}A - T^{-(k+1)}A) = 0.$$

(b) Any interval of length less than 1 is a nonrecurrent wandering set.

(c) Let A be a wandering set. Since

$$\sum_{n=0}^{\infty} \mu(A) = \sum_{n=0}^{\infty} \mu(T^{-n}A) = \mu\left(\bigcup_{n=0}^{\infty} T^{-n}A\right) + \dots + \mu(X) < \infty,$$

$\mu(A)$ must be 0. Thus T is conservative, hence infinitely recurrent.

Section A.3

1\. (a) (i) implies (ii): By hypothesis, U has a left and a right inverse; hence U is one to one and onto. Also, $\langle Uf, Ug\rangle = \langle U^*Uf, g\rangle = \langle f, g\rangle$.

(ii) implies (iii): Take $g = f$.

(iii) implies (ii): Use the polarization identity (A.2.7).

(ii) implies (i): We have $\langle U^*Uf, g\rangle = \langle Uf, Ug\rangle = \langle f, g\rangle$. Since f and g are arbitrary, $U^*U = I$. But then $U(U^*U) = U$, so if U is onto, then $UU^* = I$.

(b) This is done exactly as in (a).

(c) If $U^*g = f$, then $UU^*g = Uf$, hence, by (b), $g = Uf$. Conversely, if $Uf = g$, then

$$\|U^*g - f\|^2 = \|U^*g\|^2 - \langle U^*g, f\rangle - \langle f, U^*g\rangle + \|f\|^2$$
$$= 2\|f\|^2 - \langle g, Uf\rangle - \langle Uf, g\rangle$$
$$= 2\|f\|^2 - 2\langle f, f\rangle \quad \text{by hypothesis}$$
$$= 0.$$

(b) If $f_n \in \mathcal{F}$, $f_n \to f$ pointwise,

$$\|A_n f - A_m f\| \le \|A_n f - A_n f_k\| + \|A_n f_k - A_m f_k\| + \|A_m f_k - A_m f\|$$

$$\le 2\|f - f_k\| + \|A_n f_k - A_m f_k\|.$$

and it follows that $\{A_n f\}$ is a Cauchy sequence so that $f \in \mathcal{F}$, proving $\mathcal{F}$ closed. It is immediate that $\mathcal{F}$ is a subspace.

(c) If $f \in \mathcal{M}_1$ then $A_n f = f$ so $f \in \mathcal{F}$. If $f = g - Tg$ for some $g \in \mathcal{H}$, then $A_n f = g - T^{n+1} g$, hence $\|A_n f\| \le 2\|g\| \to 0$.

(d) If $h = Pf$, then

$$h \in \mathcal{N} \quad \text{iff} \quad h \in \mathcal{M}_1$$

$$\text{iff} \quad \langle h, g - Ug \rangle = 0 \quad \text{for all} \quad g \in \mathcal{H}$$

$$\text{iff} \quad \langle h, P(I - U)g \rangle = 0 \quad \text{for all} \quad g \in \mathcal{H}$$

$$\text{iff} \quad U^* h = h$$

$$\text{iff} \quad Uh = h \quad \text{(by (c))}$$

$$\text{iff} \quad h \in \mathcal{M}.$$

Thus $\mathcal{M} = \mathcal{N}$, and the mean ergodic theorem follows.

(e) Since $E = \lim_n A_n$ converges to a limit $\hat{E}$. Write $f = f_1 + f_2$ where $f_1 \in \mathcal{M}$, $f_2 \in \mathcal{N}$. Now $A_n f_1 \to f_1$ for all f, and also $A_n f_2 \to 0$. Therefore $g \in \mathcal{M}$ such that $f - g \perp \mathcal{M}$ then

$$\|A_n f\| = \|A_n f_1 - g\| + \|A_n g\| \le \|f_1 - g\| + \|A_n f_2\| + \|A_n g\|.$$

By (b) $A_n g \to 0$, so $\lim_n \|A_n f\| = \|f_1 - g\|$. Therefore $A_n f \to f_1 = Ef$.

(f) By definition of $\hat{S}$ and T, we have $\hat{S}T = T\hat{S} = I$. By 3.8.3, $\hat{S}$ and T are isometries and, by an invariance theorem, they must be unitary operators. By (e), if $U = T$, then $\hat{E} = E$.

6 If $\mathcal{F}$ is not an extreme point, so that a representation of the given convex positive, then $\mathcal{F}$ is preserved by T, $P_1 \le P$, and $P_1 \ne P$ (if $P_1 = P$, then $I = \lambda P_1 + \lambda P_2$ so that P_1 and P_2 being probability measures must be identical). By 5.8.10, P is not ergodic.

Conversely, assume that P is not ergodic. If A is an invariant set with $0 < P(A) < 1$, then for each $B \in \mathcal{B}$,

$$P(A \cap B) = P(A)P(A \cap B) + P(A)P(B \mid A^c)$$

$$= \lambda P(A \cap B) + \lambda P(B),$$

By the end of the proof of 5.8.10, P_1 and P_2 are preserved by T. Since $P_1(A) = 1$, if $P_1 = P$, then $P = P_2$, but $P_2(A) = P(A \mid A) = 1 \ne P(A)$. Therefore $P_1 \ne P$, so that P is not an extreme point.

7. (a) Applying the Cauchy-Schwarz inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)(Y_i-\bar{Y}_n)\right| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y}_n)^2}.$$

(b) Another way of writing the inequality in (a) is: $\hat{\rho}_{XY}^2 \le 1$. Obvious.

(c) Another way: Since $\hat{\rho}_{XY}^2 \to \rho_{XY}^2$ almost surely, hence in probability, by (a) $\hat{\rho}_{XY}^2 \le 1$ a.s.

(d) Another way: Since $\hat{\rho}_{XY}^2 \to \rho_{XY}^2$ a.s., we may multiply by I_A and integrate to obtain, by the dominated convergence theorem,

$$\int_A \hat{\rho}_{XY}^2\,dP \to \int_A \rho_{XY}^2\,dP = P(A)\rho_{XY}^2.$$

But if $\mathcal{G}$ is the σ-field of atoms $A_1, A_2, \ldots$, then

$$\begin{aligned} E(\hat{\rho}_{XY}^2 \mid \mathcal{G}) &= E(|E(\hat{\rho}_{XY}\hat{\rho}_{XY} \mid \mathcal{G})|^2) \\ &= E(\hat{\rho}_{XY}E(\hat{\rho}_{XY} \mid \mathcal{G})^*) \qquad \text{by 8.3.4} \\ &= E(\hat{\rho}_{XY}\hat{\rho}_{XY}^*) \qquad \text{by 8.3.3}. \end{aligned}$$

Under the hypothesis, $\hat{\rho}_{XY} \to \rho(X,Y)$ a.s., $\hat{\rho}_{XY}^* \to \rho(\bar{X},\bar{Y})$ a.s., proving (d).

(e) Applying (c) to $I_A, I_B \in \mathcal{F}$ and $\varepsilon > 0$, choose $A_1, B_1 \in \mathcal{F}_1$ such that $P(A \triangle A_1) < P(A)\varepsilon$, $P(B \triangle B_1) < P(B)\varepsilon$, and $\rho(I_A, I_B) \le \rho(I_{A_1}, I_{B_1}) + 2\varepsilon$ (see the proof of 8.2.7). Then

$$\begin{aligned} &\left|\frac{1}{n}\sum_{i=1}^{n} P(A\cap T^{-i}B) - P(A)P(B)\right| \\ &\quad \le \left|\frac{1}{n}\sum_{i=1}^{n-1}[P(A_1\cap T^{-i}B_1) - P(A_1)P(B_1)]\right| \\ &\quad + \left|\frac{1}{n}\sum_{i=1}^{n-1} P(A_1\cap T^{-i}B_1) - P(A)P(B)\right| \\ &\quad + P(A_1\cap T^{-n}B_1) + P(A_1)P(B_1). \end{aligned}$$

The last term is less than ε for all n, and the second is less than ε for large n by (c). Since the third term is less than 2ε, the result follows.

(f) Another way: Let $\mathcal{B}$ be the (countable) set consisting of $\emptyset$, Ω, and all finite unions of sets of the form $A_i \cap T^{-j}A_k$. Then $P(A \cap T^{-n}B) \to P(A)P(B)$ for all $A, B \in \mathcal{B}$. By (c), $P(A \cap T^{-n}B) \to P(A)P(B)$ for all $A, B \in \mathcal{F}$. Now let $\mathcal{F}$ be the class of all A such that $P(A \cap T^{-n}B) \to P(A)P(B)$ for all $B \in \mathcal{F}$. Clearly $\mathcal{F}$ is a class

Section 8.1

8. (a) We have $E(B_n) = \sum_{k=1}^{n} P(B_k)$, and

$$P(B_k) = P(X_k \neq 0, X_k \neq X_{k-1}, X_{k-1} \neq X_{k-2}, \ldots, X_2 \neq X_1) = P(X_1 \neq 0) P(X_2 \neq X_1 \mid X_1 \neq 0) \cdots P(X_k \neq X_{k-1} \mid X_{k-1} \neq X_{k-2}, \ldots, X_2 \neq X_1),$$

since the X_i are independent.

There are 2^k k-digit binary numbers, and $P(B_k)$ is the arithmetic average of a sequence consisting of $P(A)$ followed by $2^k - 1$ terms each equal to $P(A^c)$.

(b) Let $Z_k = X_k - X_{k-1}$ for $k \geq 2$. Then $Z_2, Z_3, \ldots$ are independent with $P(Z_k = 1) = P(Z_k = -1) = \frac{1}{2}$, and $X_n = X_1 + Z_2 + \cdots + Z_n$. Therefore, the strong law of large numbers yields the desired result.

(c) For any positive integer n, $\{2^{k-1} + 1 \leq n \leq 2^k\}$ for some k. Hence $M_n = X_1 X_2 \cdots X_n$ is nonzero only if $X_1 = X_2 = \cdots = X_{2^{k-1}}$. Therefore,

$$\frac{M_n}{n} \leq \frac{X_1 X_2 \cdots X_{2^{k-1}}}{2^{k-1}} = \frac{1}{n}.$$

Since $\lim_{n \to \infty} \frac{1}{n} = 0$, it follows that $\lim_{n \to \infty} \frac{M_n}{n} = 0$ with probability 1 by the definition of almost sure convergence. Now $Z_n = \frac{M_n}{n}$ and M is arbitrary, so the result follows from (b).

(d) Since $p_{ij} = 0$ if $X_{k-1} \neq i$ or $X_k - X_{k-1} \neq j - i$, and $p_{ij} = \frac{1}{2}$ otherwise, P_k can be expressed as $\phi(X_{k-1}, X_{k-2}, \ldots)$ where $\phi: M^{k-1} \to M^k$. Moreover, relative to $\{\phi(\vec{x})\}^{k-1}$ the primitive events are equally likely.

(e) The sum $\sum_{k=1}^{n} V_k$ is the number of steps taken by the random walk before the walk has visited all n points. If $i \neq j$ and X_k and X_l are never visited on the same step, then $\sum_{k=1}^{n} V_k \neq n$. By (c), $\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} V_k = E(V_1) \neq 0$. But since the X_k are all equal,

$$E(V_1) = P(X_2 \neq X_1, X_3 \neq X_1, X_3 \neq X_2, \ldots, X_n \neq X_1) = 0.$$

Section 8.3

1. Let $\{X_n\}$ and $\{X_n'\}$ be two-state ergodic sequences with alphabet M and M'. Flip a coin. If the result is heads, let $Z_n = X_n$ for all n, and if tails, let $Z_n = X_n'$ for all n. If p is the probability of heads, the limit relative frequency of any given probability p and M' with probability $1 - p$, so that the entropy of $\{Z_n\}$ is $p\phi(M) + (1-p)\phi(M')$. In part (a), choose $M = M'$, and in part (b), choose $M \neq M'$. There is no problem in realizing these choices; for example, if the Z_n are independent and take on $+1$ or -1 with equal probability, then $\phi = \log 2$.

Index

accounting procedures, 112, 113, 115, 116, 117
acquisitions, 104, 105
Adams-Weiss, Jessica, 245
advice, 112, 114

Air Transport Association
  International Air Transport Association, 369
Alcoholics Anonymous, 210
  attendance, 211
  effectiveness, 211
  effectiveness, 88
Allen, Christopher, 22
Allen, Christopher, senior partner, 23–24
Allen, Jessica
  attendance, 75
  client interviews, 74–78
Allen, Jessica
  probability, opinion, 157
  research, 111
  results, 177
Allen, Robert, Jessica, 22, 23, 30, 31, 32, 37
  attendance, 11
  attendance, 76
Allen, Robert, Jessica, and Associates, 15, 17
  Internal Revenue Service, 90, 102, 103
Allen, Robert, 122
  family, 122, 127
Allen, Robert, Jessica, and Associates, 15, 17
  Internal Revenue Service, 90, 102, 103
Allen, Robert, 122
  family, 122, 127

Appeals, Court of, 33
Arbitration, Internal Revenue Service, 178
arbitration, appeals, 200
arrest, 111, 113, 115, 116
Arthur, Christopher, 129–130, 131, 132
Arthur, Christopher, 129–130, 131, 132
Arthur, Christopher, 129–130, 131, 132

Radiocarbon dating, 214–15
Rainwater harvesting, 290

Salt, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135,

poetic epithets of Athena 100, 102, 124, 126, 128
poetry
  of Aristophanes 11, 181
  epic 148, 184
  lyric 202
polytheism, development towards 422
portraiture 324, 334
  of Augustus 348, 350, 351
  Hellenistic 332
Praxiteles 310, 311
priesthoods 57
Ptolemaic dynasty 57

quadrilateral arrangement of Roman temples 408–410
quadrivium 91

Hellenistic 68
Roman 70, 71
  decline of 69
  history 68, 69
realism
  Attic 78
  Egyptian 224, 241
  Hellenistic 311
  Roman 315
  Greek 176, 178
reasoning 132, 133
reconstruction 167, 170, 190, 192
rectangular temple plan 398
religious beliefs 131
religion
  Christian 190, 191, 192, 193
  Egyptian 144, 145
  Greek
    Archaic 48
    Hellenistic 418
    Roman 47, 49
Renaissance 25
restoration 248, 250

sacrifice
  animal 54
  human 54
  Roman 54
Sappho 98, 99
Sarcophagus of Lucius Cornelius Scipio Barbatus 271
Satyrus 332
schools 68, 69
sculpture, Greek 95
  Hellenistic 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348,

[illegible]

ISBN 0-12-063202-1

ISBN 0-12-065202-1